5-2015

A Statistical Distance Approach to Dissimilarities in Ecological Data

Dominique Jerrod Morgan
Clemson University

A Statistical Distance Approach to Dissimilarities in Ecological Data

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
Dominique Jerrod Morgan
May 2015

Accepted by:
Dr. Chanseok Park, Committee Chair
Dr. William Bridges
Dr. Calvin Williams
Dr. Chris McMahan

Abstract

Statistical distances allow us to quantify the *closeness* between two statistical objects. Many distances are important in statistical inference, but can also be used in a wide variety of applications through goodness-of-fit procedures. This dissertation aims to develop useful theory and applications for these types of procedures.

Extensive research has already been done for statistical distances in parameter estimation and hypothesis testing. Specifically, there are a large number of distances that can be chosen to minimize the difference between the data and the assumed model. This procedure is known as the minimum distance approach. It was not necessary that the statistical distance be symmetric in parameter estimation but there are many applications in goodness-of-fit testing that require symmetric distance functions. In this paper, one of the main goals is to establish theory for selecting an appropriate symmetric distance when being used in these types of applications. Secondly, we propose a new class of symmetric distances that share the same desirable properties as previously proposed methods.

In addition to focusing on symmetric statistical distances, a new method will be proposed for determining whether or not a particular distance is efficient or robust. Lastly, we exhibit the usefulness of our approach through applications in ecology and image processing.

Dedication

This work is dedicated to my wonderful family and friends. In particular, my awesome parents and brothers. Without your words of encouragement and positive energy, I do not know where I would be. Thank you for everything!

Acknowledgments

I would like to express my gratitude to many individuals who helped me during my time in Clemson. Special thanks to my advisor, Dr. Chanseok Park, for his guidance, support, and patience. I also would like to thank my committee members for their willingness to help me in this process. Lastly, I would like to thank all staff and fellow graduate students who made my experience at Clemson a great one.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Distances in Mathematics

Distance measures are used to determine how close or far away two objects are in space. In mathematics, this distance typically assumes a certain structure and behaves according to a specific set of rules. For example, for any f, g, $h \in X$, the metric is a distance function $\text{Dist} \to \mathbb{R}$ defined on $X \times X$ that satisfies the following properties:

1. $\text{Dist}(f, g) \geq 0$ (non-negative)
2. $\text{Dist}(f, g) = 0$ if and only if $f = g$ (identity)
3. $\text{Dist}(f, g) = \text{Dist}(f, g)$ (symmetry)
4. $\text{Dist}(f, g) \leq \text{Dist}(f, h) + \text{Dist}(h, g)$ (triangle inequality)

The first two properties are necessary for any distance, but the third and fourth properties allow us to have really nice geometrical interpretations for the distance. One of the most popular and fundamental metrics in mathematics is the Euclidean

distance. For any f, $g \in X$, this metric represents the length of the line connecting f and g. The Euclidean distance is extremely useful when you simply would like to know the shortest geographical distance between two points, however the usefulness of this particular distance depends on the circumstances. What if we would like to classify two objects (statistically or ecologically) as either being *too close* or *too far away*? Once the Euclidean distance is calculated, there is no way to answer such a question. Therefore, it is necessary to look at different types of distance functions.

1.2 Distances in Statistics

To determine if the distance between two objects is statistically meaningful, it makes sense to consider distance functions that follow known probability distributions. Although the geometric interpretations of a metric are still required in some cases, many distances in statistics do not satisfy the conditions of a metric. Specifically, the symmetric property and triangle inequality are often unnecessary for some of the most popular statistical distances.[1]

Statistical distances are a fundamental part of statistical inference. In parametric statistical inference, a model is usually selected that will minimize (or maximize) an appropriate statistical distance that measures how close the data are to the hypothesized model. Furthermore, in goodness-of-fit testing, a statistical distance can be used to determine whether the data and model are *close enough* to each other based on some specified criteria. As you can imagine, there are many reasonable distances one could define to quantify the closeness between the data and model. One of the most popular distances of this type is the log-likelihood ratio statistic, which is perhaps one of the most standard procedures used for parameter estimation.

[1]There are some that refer to such a distance as a *divergence* or a *deviance*.

The popularity of the log-likelihood ratio statistic is in large part due to the fact that it yields an estimate, called the maximum likelihood estimate, that is superior in efficiency under certain model conditions. However, there are still improvements to be made in this area, and thus a reason to explore other statistical distances. For instance, the maximum likelihood estimate is inefficient when there are deviations from the assumed underlying model or in the presence of outliers or contamination.

In the last sixty years, extensive research has been done to explore a wide variety of alternative distances that can yield estimates that can remain efficient when model conditions are violated. In [Cressie and Read, 1984], a large family of popular distances known as the power divergence family was developed. A decade later, an even more general form of distances was introduced in [Lindsay, 1994]. Following this approach, many distances were able to be defined after that which led to a comprehensive overview being given recently in [Basu et al., 2011].

A statistical procedure is considered to be robust if the result is relatively insensitive to small changes in the underlying model, small changes in the bulk of the observations (*inliers*), or large changes in a small number of observations (*outliers*) [Jurekov and Sen, 1996]. Typically, outliers are due to either measurement error or data being drawn from a different distribution. Some robust procedures perform well even when as much as half of the data are outliers. Robust procedures are especially useful in statistical inference [Basu et al., 2011].

Similarly, this paper will refer to a statistical procedure as inlier robust if the output is relatively insensitive to small changes in the bulk of the observations. There are also instances where an inlier robust procedure, one that is sensitive to outliers and relatively insensitive to inliers, is preferred. For example, in goodness of fit testing, where it is determined whether two statistical objects are *close enough*, an inlier robust procedure will be much more appropriate than a robust one in this

particular case [Basu et al., 2011].

1.3 Research Questions

This paper aims to accomplish two main goals. The first goal is to derive a new class of statistical distances that can be used to solve a wide range of problems. A good number of these problems require a distance measure that is symmetric. Since this property is not necessary in parameter estimation and hypothesis testing, there is a limited number of symmetric statistical distances. Later in this paper, a new class of symmetric statistical distances will be introduced and the advantages will be discussed in more detail. These new symmetric distances will eventually be used in a few interesting applications, including a problem in ecology. Of particular interest will be choosing an appropriate distance function that describes the measure of dissimilarity between two ecological sites. The distance between two sites is calculated by comparing the relative species abundance at each site. Another application that will be considered is in image segmentation, the process of partitioning an image so that it is represented in a more meaningful manner. One method of performing this kind of partition involves measuring the dissimilarity individual pixels have to "neighbors."

The second goal of this paper is to improve methods of choosing a statistical distance. Specifically, more insight about choosing the most appropriate distance to use in a specific situation would be very helpful. For example, a modified procedure that can be used to quickly determine the robustness and efficiency properties of a statistical distance will be proposed. In [Lindsay, 1994], a residual adjustment function was defined and it was determined that the desirable properties looked for in a statistical distance could be explained by this function. The problem with using

this function is that it is defined on an infinite domain and there is no meaningful graphical representation. A weighted likelihood function was then proposed in [Park et al., 2002], but it is our belief that the update made to this weight function will make interpretation of desirable properties easier to understand. An entirely new weight function that will make it easier to choose the most appropriate symmetric distance function when working in problems outside of parameter estimation will be introduced. To the best of our knowledge, this kind of a symmetric statistical distance approach has not been done before.

Chapter 2

Statistical Distances

In this chapter, some popular statistical distances will be introduced, the asymptotic distribution of each will be investigated, and then methods for comparing distances will be discussed. For simplicity, discrete random variables are considered but it should be noted that our discussion can be generalized for continuous random variables as well.

2.1 Disparity-based Distances

Suppose that Y represents a discrete random variable with probability mass function (or pmf) f_θ and y is its realization. Let d be the empirical probability mass function (relative frequency).

Typically of interest is finding an estimator, say $\hat{\theta}$, for the true unknown parameter θ. The statistical distance approach to parameter estimation is to find the value of θ which will minimize a specified distance between the model f_θ and the empirical probability mass function d.

One classical parameter estimation method that is known to produce an effi-

cient estimator under regular conditions is the maximum likelihood approach. The earliest works of the maximum likelihood approach can be traced back to the 1700s [Stigler, 2007], but the approach as it is known today was developed in [Fisher, 1922]. In general, for a fixed data set and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes the "agreement" of the selected model with the observed data. Given observations $y_1, \ldots, y_n$, the likelihood function is given by

$$L = \prod_{i=1}^{n} f_\theta(y_i)$$

Maximizing the likelihood is equivalent to maximizing the log-likelihood function

$$\ell = \sum_{i=1}^{n} \log f_\theta(y_i).$$

Therefore the maximum likelihood estimate can be written as

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta L = \arg\max_\theta \prod_{i=1}^{n} f_\theta(y_i),$$

which is equivalent to

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \ell = \arg\max_\theta \sum_{i=1}^{n} \log f_\theta(y_i).$$

Consider rewriting the values $y_1, \ldots, y_n$ as k distinct values, say $y_1^*, \ldots, y_k^*$. For $j = 1, \ldots, k$, let n_j represent the frequency for the value y_j^*. Note that $\sum_{j=1}^{k} n_j = n$. Let $d(y_j^*) = \frac{n_j}{n}$ for $j = 1, \ldots, k$ represent the relative frequencies for each y_j^*. For example, suppose we have $(y_1, y_2, y_3, y_4, y_5) = (1.2, 2.1, 1.2, 2.1, 5.4)$. Then $(y_1^*, y_2^*, y_3^*) = (1.2, 2.1, 5.4)$, $(n_1, n_2, n_3) = (2, 2, 1)$, and $(d(y_1^*), d(y_2^*), d(y_3^*)) = (0.4, 0.4, 0.2)$. So the MLE can be rewritten as

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\max_{\theta} \sum_{j=1}^{k} n_j \log f_\theta(y_j^*) \\ &= \arg\max_{\theta} \sum_{j=1}^{k} n d(y_j^*) \log f_\theta(y_j^*)\end{aligned}$$

Without loss of generality we will let $y_1 = y_1^*, \ldots, y_k = y_k^*$. Therefore the MLE can be expressed as

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\max_{\theta} \sum_{i=1}^{k} n d(y_i) \log f_\theta(y_i) \qquad (2.1) \\ &= \arg\min_{\theta} \sum_{i=1}^{k} (-2n d(y_i) \log f_\theta(y_i)) \\ &= \arg\min_{\theta} \left[2n \sum_{i=1}^{k} d(y_i) \log d(y_i) - 2n \sum_{i=1}^{k} d(y_i) \log f_\theta(y_i) \right] \\ &= \arg\min_{\theta} 2n \left[\sum_{i=1}^{k} d(y_i) \log d(y_i) - \sum_{i=1}^{k} d(y_i) \log f_\theta(y_i) \right] \\ &= \arg\min_{\theta} 2n \sum_{i=1}^{k} \left[d(y_i) \log \frac{d(y_i)}{f_\theta(y_i)} \right]. \qquad (2.2)\end{aligned}$$

So this means that maximizing the likelihood function is equivalent to minimizing the distance

$$LD(d, f_\theta) = 2n \sum_{i=1}^{k} \left[d(y_i) \log \frac{d(y_i)}{f(y_i)} \right].$$

This is closely related to the log-likelihood ratio statistic which is used in goodness-of-fit testing

$$G^2 = 2 \sum_{i=1}^{k} \left[O_i \log \frac{O_i}{E_i} \right]$$

where $E_i = nf(y_i)$ and $O_i = nd(y_i)$ represent the expected and observed frequencies of the value y_i.

Another popular distance used in goodness-of-fit testing is the Pearson's chi-square statistic introduced by [Pearson, 1900] as

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

We will show that X^2 has a relationship to G^2. Consider the Pearson residual which is defined by [Pearson, 1900] as

$$\delta_P(x) = \frac{d(x) - f_\theta(x)}{f_\theta(x)} = \frac{d(x)}{f_\theta(x)} - 1$$

Here δ_P is known as Pearson's residual. By definition, we have $\delta_P(x) \geq -1$ for any value of x. We can rewrite the Pearson's chi-square (PCS) goodness-of-fit statistic using this residual $\delta_P(x)$ as below:

$$\begin{aligned}
\text{PCS} &= n \sum_{i=1}^{k} \left[\frac{(d(y_i) - f_\theta(y_i)}{f_\theta(y_i)} \right]^2 f_\theta(y_i) \\
&= n \sum_{i=1}^{k} \delta_P^2(y_i) f_\theta(y_i) \\
&= nE\left[\delta_P^2(Y)\right] \qquad (2.3)
\end{aligned}$$

This statistical divergence was first introduced in [Pearson, 1900] and is still very popular, in large part, due to the fact that it approximates an chi-square distribution. This means that once we are able to assess a statistical significance of a measurement since it follows a known distribution.

Using the Pearson's residual ($\delta_P(x) = \frac{d(x)}{f(x)} - 1$), it can be seen how the log-likelihood ratio statistic in equations (2.1) and (2.2) can be rewritten as

$$\begin{aligned}
\text{LD} &= 2n \sum_{i=1}^{k} \left[d(y_i) \log \frac{d(y_i)}{f_\theta(y_i)} \right] \\
&= 2n \sum_{i=1}^{k} \left[(\delta_P(y_i) + 1) \log(\delta_P(y_i) + 1) \right] f_\theta(y_i) \\
&= 2nE\left[(\delta_P(Y) + 1) \log(\delta_P(Y) + 1) \right] \qquad (2.4)
\end{aligned}$$

Using the Taylor series expansion at $\delta_P = 0$, we get $(\delta_P + 1)\log(\delta_P + 1) = \delta_P + \frac{\delta_P^2}{2} + o(\delta_P^2)$. Applying this to equation (2.4) and using equation (2.3), the following result is obtained in (2.5).

$$\begin{aligned} \text{LD} &= 2nE\left[(\delta_P(Y)+1)\log(\delta_P(Y)+1)\right] \\ &\approx 2nE\left[\delta_P(Y)+\frac{1}{2}\delta_P^2(Y)\right] \\ &= nE\left[\delta_P^2(Y)\right] = \text{PCS} \end{aligned} \tag{2.5}$$

since $E\left[\delta_P(Y)\right] = E\left[\frac{d(y)-f_\theta(y)}{f_\theta(y)}\right] = \sum_{\forall y}\Big(d(y)-f_\theta(y)\Big) = 1-1 = 0.$

This result means that in goodness of fit testing, the log-likelihood ratio statistic G^2 can be approximated by the Pearson's chi-square statistic X^2. Therefore we can perform hypothesis tests, such as the likelihood ratio test (LRT). For the rest of this section, we will show a similar result for other popular statistical distances.

Now consider the Neyman's chi-square statistic defined as

$$\text{NCS} = \sum_{j=1}^{k}\frac{(O_j-E_j)^2}{O_j}. \tag{2.6}$$

This distance is the same as the Pearson's chi-square statistic except for now the denominator is being divided by O_j instead of E_j. It was introduced in [Neyman, 1949]. Similar to the log-likelihood ratio statistic and the Pearson's chi-square statistic, the Neyman's chi-square statistic is also used in goodness-of-fit testing.

Using the Pearson's residual, we can see that the Neyman's chi-square statistic in equation (2.6) can be rewritten as

$$
\begin{aligned}
\text{NCS} &= n \sum_{i=1}^{k} \frac{(d(y_i) - f_\theta(y_i))^2}{d(y_i)} \\
&= n \sum_{i=1}^{k} \Big[\frac{(d(y_i) - f_\theta(y_i))^2}{(\delta_P(y_i) + 1) f_\theta(y_i)} \Big] \\
&= n \sum_{i=1}^{k} \frac{1}{\delta_P(y_i) + 1} \left[\frac{d(y_i) - f_\theta(y_i)}{f_\theta(y_i)} \right]^2 f_\theta(y_i) \\
&= n \sum_{i=1}^{k} \Big[\frac{\delta_P^2(y_i)}{\delta_P(y_i) + 1} \Big] f_\theta(y_i) \\
&= nE \Big[\frac{\delta_P^2(Y)}{\delta_P(Y) + 1} \Big]
\end{aligned} \tag{2.7}
$$

Also using the Taylor series expansion at $\delta_P = 0$, we get $\frac{\delta_P^2}{\delta_P + 1} = \delta_P^2 + o(\delta_P^2)$. Combining this with equations (2.3) and (2.7), we obtain the following result

$$
\text{NCS} = nE \Big[\frac{\delta_P^2(Y)}{\delta_P(Y) + 1} \Big] \approx nE \left[\delta_P^2(Y) \right] = \text{PCS}. \tag{2.8}
$$

So similar to the likelihood ratio statistic, the Neyman's chi-square statistic can be approximated by the Pearson's chi-square statistic.

In this paper we will consider variations of the chi-square distances. The symmetric chi-square statistic is defined as

$$
\text{SCS} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{(O_j + E_j)}. \tag{2.9}
$$

A generalized version of the Pearson and Neyman chi-square statistics, called the blended weight chi-square statistic, proposed by [Lindsay, 1994], will also be consid-

ered. The blended weight chi-square statistic is defined as

$$\text{BWCS}_\tau = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{\Big[\tau O_j + (1-\tau)E_j\Big]} \tag{2.10}$$

where $\tau \in [0, 1]$. Note that when $\tau = 0$, the distance in equation (2.10) becomes the Pearson's chi-square statistic. When $\tau = 1$, the distance in equation (2.10) becomes the Neyman's chi-square statistic. It can be shown that both the symmetric and blended weight chi-square statistics can also be approximated by the Pearson's chi-square statistic in (2.3).

Another popular measure of discrepancy in statistical inference is the squared Hellinger distance

$$H^2 = 2\sum_{j=1}^{k}(\sqrt{O_j} - \sqrt{E_j})^2. \tag{2.11}$$

Although the Hellinger distance has been around for many years, the minimum Hellinger distance approach was explored in [Beran, 1977]. Unlike the previous distances we have considered so far, this distance satisfies the axioms of a metric. So it would be preferred in problems when a geometric interpretation is required. The squared Hellinger distance in equation (2.11) can be rewritten as

$$
\begin{aligned}
H^2 &= 2\sum_{i=1}^{k}(\sqrt{nd(y_i)} - \sqrt{nf_\theta(y_i)})^2 \\
&= 2n\sum_{i=1}^{k}\left[d(y_i) + f_\theta(y_i) - 2\sqrt{f_\theta(y_i)d(y_i)}\right] \\
&= 2n\sum_{i=1}^{k}\left[\frac{d(y_i)}{f_\theta(y_i)} + 1 - 2\sqrt{\frac{d(y_i)}{f_\theta(y_i)}}\right] f_\theta(y_i) \\
&= 2n\sum_{i=1}^{k}\left[\delta_P(y_i) + 2 - 2\sqrt{\delta_P(y_i)+1}\right] f_\theta(y_i) \\
&= nE\Big[2\delta_P(Y) + 4 - 4\sqrt{\delta_P(Y)+1}\Big] \qquad (2.12)
\end{aligned}
$$

Using the Taylor series expansion at $\delta_P = 0$, we get $2\delta_P + 4 - 4\sqrt{\delta_P + 1} = \delta_P^2 + o(\delta_P^2)$. Combining this with equations (2.3) and (2.12), the following result is obtained

$$
H^2 = nE\left[2\delta_P(Y) + 4 - 4\sqrt{\delta_P(Y)+1}\right] \approx nE\left[\delta_P^2(Y)\right] = \text{PCS}. \qquad (2.13)
$$

Again, this result implies that the squared Hellinger distance can be approximated by a Pearson's chi-square statistic in (2.3). A generalized version of the squared Hellinger distance is called the blended weight Hellinger distance, proposed by [Lindsay, 1994], which is defined as

$$
\text{BWHD}_\tau = \sum_{i=1}^{k}\left[\frac{d(y_i) - f_\theta(y_i)}{\tau\sqrt{d(y_i)} + (1-\tau)\sqrt{f_\theta(y_i)}}\right]^2 \qquad (2.14)
$$

where $\tau \in [0, 1]$. Note that when $\tau = 0$, the distance in equation (2.14) becomes the Pearson's chi-square statistic. When $\tau = 0.5$, the distance in equation (2.14) becomes

the squared Hellinger distance. When $\tau = 1$, the distance in equation (2.14) becomes the Neyman's chi-square statistic. It is our belief that the blended weight Hellinger distance can also be approximated by the Pearson's chi-square statistic in (2.3).

Lastly we will consider common variations of the likelihood ratio test statistic. By swapping the O_j's with the E_j's in equation (2.2), we obtain the Kullback-Leibler distance

$$\mathrm{KL} = 2\sum_{j=1}^{k} E_j \log \frac{E_j}{O_j}. \tag{2.15}$$

The generalized Kullback-Leibler distance, a combination between the likelihood ratio statistic and the Kullback-Leibler distance, is defined as

$$\mathrm{GKL}_\tau = 2\sum_{j=1}^{k} \left[\frac{O_j}{1-\tau} \log\left(\frac{O_j}{E_j}\right) - \left(\frac{O_j}{1-\tau} + \frac{E_j}{\tau}\right) \log\left(\tau\frac{O_j}{E_j} + 1 - \tau\right) \right] \tag{2.16}$$

where $\tau \in (0, 1]$ [Park and Basu, 2003]. It is derived from mixing the log-likelihood ratio statistic (GKL with $\tau = 0$) and the Kullback-Leibler (KL) distance (GKL with $\tau = 1$). Using equation (2.4), we can see that

$$\mathrm{GKL}_\tau = \frac{2n}{1-\tau} E\Big[\Big(\delta_P(Y) + 1\Big) \log\Big(\delta_P(Y) + 1\Big)\Big] - c$$

where

$$
\begin{aligned}
c &= 2n \sum_{i=1}^{k} \Big[\frac{(d(y_i) - f_\theta(y_i))\tau + f_\theta(y_i)}{\tau(1-\tau)} \log\Big(\tau(\delta_P(y_i)+1) + 1 - \tau\Big)\Big] \\
&= 2n \sum_{i=1}^{k} \frac{\delta_P(y_i) f_\theta(y_i)\tau + f_\theta(y_i)}{\tau(1-\tau)} \log(\tau\delta_P(y_i)+1) \\
&= 2n \sum_{i=1}^{k} \Big[\frac{\tau\delta_P(y_i)+1}{\tau(1-\tau)} \log(\tau\delta_P(y_i)+1)\Big] f_\theta(y_i) \qquad (2.17)
\end{aligned}
$$

Similar to the log-likelihood ratio statistic, we can see that using the Taylor series expansion about $\delta_P = 0$ we have $(\tau\delta_P + 1)\log(\tau\delta_P + 1) = \tau\delta_P + \frac{\tau^2}{2}\delta_P^2 + o(\delta_P^2)$. Combining this approximation with equations (2.3) and (2.17), we obtain the following result.

$$
\begin{aligned}
\mathrm{GKL}_\tau &\approx \frac{2n}{1-\tau} E\left[\delta_P(Y) + \frac{\delta_P^2(Y)}{2}\right] - \frac{2n}{\tau(1-\tau)} E\left[\tau\delta_P(Y) + \frac{\tau^2}{2}\delta_P^2(Y)\right] \\
&= nE\left[\frac{1}{1-\tau}\delta_P^2(Y) - \frac{\tau}{1-\tau}\delta_P^2(Y)\right] \\
&= nE\left[\delta_P^2\right] \\
&= \mathrm{PCS}. \qquad (2.18)
\end{aligned}
$$

So the generalized Kullback-Leibler distance can also be approximated by the Pearson's chi-square statistic in (2.3).

Figure 2.1 summarizes the results of the approximations of the distances above. So far we have looked at several types of popular statistical distances and shown that all of them can be approximated by the Pearson's chi-square statistic in (2.3). Why is this important? Our results allow these statistical distances to be used for inference and, in particular, hypothesis testing. For example, the likelihood ratio test

(LRT) can decide whether or not the data follows a particular model with density $f_\theta(y)$. Equivalently, we can use any of the distances in this chapter to perform the asymptotically equivalent test since they all have the same asymptotic distribution.

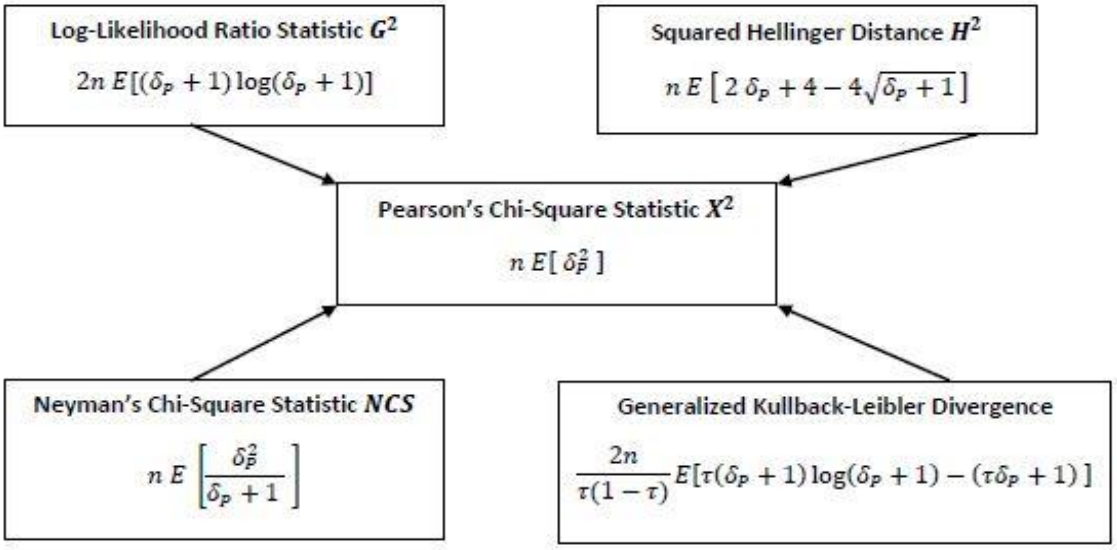

Figure 2.1: Chi Square Approximations Using Taylor Series

It is worth mentioning that we have only highlighted two symmetric distances so far: the symmetric chi-square statistic and the squared Hellinger distance. It is also easy to see that $\mathrm{GKL}_{\tau=0.5}$ is also symmetric. Although there are several more common symmetric statistical distances, we will introduce a new class of symmetric distances later and show that they can also be approximated by a Pearson's chi-square statistic in (2.3).

2.2 Investigation of Statistical Disparities

In the previous chapter, we mentioned that the maximum likelihood estimator maximizes the "agreement" between the selected model and the observed data.

Intuitively, this is the same as minimizing the "distance" between the selected model and the observed data. The results in the previous section tell us that there are many suitable choices for distance functions we can choose to minimize. In general, this procedure is called the minimum distance approach. When using the minimum distance approach in parameter estimation, the goal is to obtain an estimator that is both efficient under model conditions and robust when there are deviations from the selected model (this can be caused by outliers, measurement error, or samples drawn from a different population). Unfortunately, there is usually a trade-off between efficiency and robustness. Therefore it is very important to have criteria to determine whether or not a statistical distance will have robust properties while remaining efficient.

A "disparity" measures the discrepancy between the empirical density $d(\cdot)$ and the model density $f_\theta(\cdot)$ based on the function $G(\cdot)$, and is defined in [Lindsay, 1994] as

$$\rho_G(d, f) = \sum_{i=1}^{k} G(\delta_P(y_i)) f_\theta(y_i) = \sum_{i=1}^{k} G\Big(\frac{d(y_i) - f_\theta(y_i)}{f_\theta(y_i)}\Big) f_\theta(y_i) \tag{2.19}$$

where G is thrice differentiable convex function on $[-1, \infty)$ with $G(0) = 0$. Assuming the model f_θ is differentiable, the minimum disparity estimator can then be found by solving an estimating equation of the form

$$-\nabla \rho_G = \sum_{\forall y} A(\delta_P(y)) \nabla f_\theta(y) = 0 \tag{2.20}$$

where $A(\delta_P) = (1 + \delta_P)G'(\delta_P) - G(\delta_P)$ is called the residual adjustment function (RAF). Typically, the RAF is centered and rescaled so that $A(0) = 0$ and $A'(0) = 1$.

The robustness and efficiency properties of the minimum disparity estimators are explained by the properties of the RAF function. For example, large outliers, corresponding to large positive values of δ_P, are much better controlled by disparities having the property $A(\delta_P)/\delta_P \to 0$ as $\delta_P \to \infty$. Efficiency measures the optimality of an estimator, where efficiency of higher orders imply that optimality can be achieved by using fewer observations. A necessary condition for second order efficiency of a minimum disparity estimator is $A''(0) = 0$ [Lindsay, 1994]. Table 2.1 shows the $G(\cdot)$ and $A(\cdot)$ functions for each distance defined in the previous section. Of particular interest, is the fact that three of these distances are actually symmetric. Recall that $\text{BWCS}_{\tau=0.5} = \text{SCS}$ and $\text{BWHD}_{\tau=0.5} = \text{HD}$.

Figures 2.2 and 2.3 show the graphs for the $G(\cdot)$ and $A(\cdot)$ functions.

Disparity	$G(\delta_P)$	$A(\delta_P)$	Symmetric?
LD	$(\delta_P+1)\log(\delta_P+1)-\delta_P$	δ_P	No
KL	$-\log(\delta_P+1)+\delta_P$	$\log(\delta_P+1)$	No
PCS	$\frac{1}{2}\delta_P^2$	$\delta_P+\frac{1}{2}\delta_P^2$	No
NCS	$\frac{\delta_P^2}{2(\delta_P+1)}$	$\frac{\delta_P}{\delta_P+1}$	No
SCS	$\frac{\delta_P^2}{2(\delta_P+2)}$	$\frac{\delta_P(3\delta_P+4)}{(\delta_P+2)^2}$	**Yes**
HD	$2(\sqrt{\delta_P+1}-1)^2$	$2(\sqrt{\delta_P+1})$	**Yes**
GKL_τ	$\frac{(\delta_P+1)}{1-\tau}\log(\delta_P+1)-\frac{(\tau\delta_P+1)}{\tau(1-\tau)}\log(\tau\delta_P+1)$	$\frac{1}{\tau}\log(\tau\delta_P+1)$	**at** $\tau=0.5$
BWCS_τ	$\frac{\delta_P^2}{2(\tau\delta_P+1)}$	$\frac{\delta_P}{\tau\delta_P+1}+\frac{1-\tau}{2}\left[\frac{\delta_P}{\tau\delta_P+1}\right]^2$	**at** $\tau=0.5$
BWHD_τ	$\frac{\delta_P^2}{2}\left[\tau\sqrt{\delta_P+1}+1-\tau\right]^2$	$\frac{\delta_P}{2}\left[\tau\sqrt{\delta_P+1}+1-\tau\right]^2+\frac{1-\tau}{2}\frac{\delta_P^2}{2}\left[\tau\sqrt{\delta_P+1}+1-\tau\right]^2$	**at** $\tau=0.5$

Table 2.1: G and A functions for popular disparities

When looking at the graph of the RAF function for a statistical distance, of particular interest are the values of $A(-1)$, $A(0)$, and $A(\infty)$. Specifically, behavior of the RAF around the point $\delta_P = 0$ determines the efficiency of an estimator. Also $A(-1)$ and $A(\infty)$ give us insight on whether the minimum disparity estimator is anti-robust or robust, respectively. For the RAF functions shown in Figures 2.2 and 2.3, it is clear that by design $A(0) = 0$ for all distances. However, the values of $A(-1)$ and $A(\infty)$ are not as clear based on the graphs. It is then necessary to consider a weight function based on the RAF function that can also be used as a graphical representation to interpret the robustness of minimum disparity estimators.

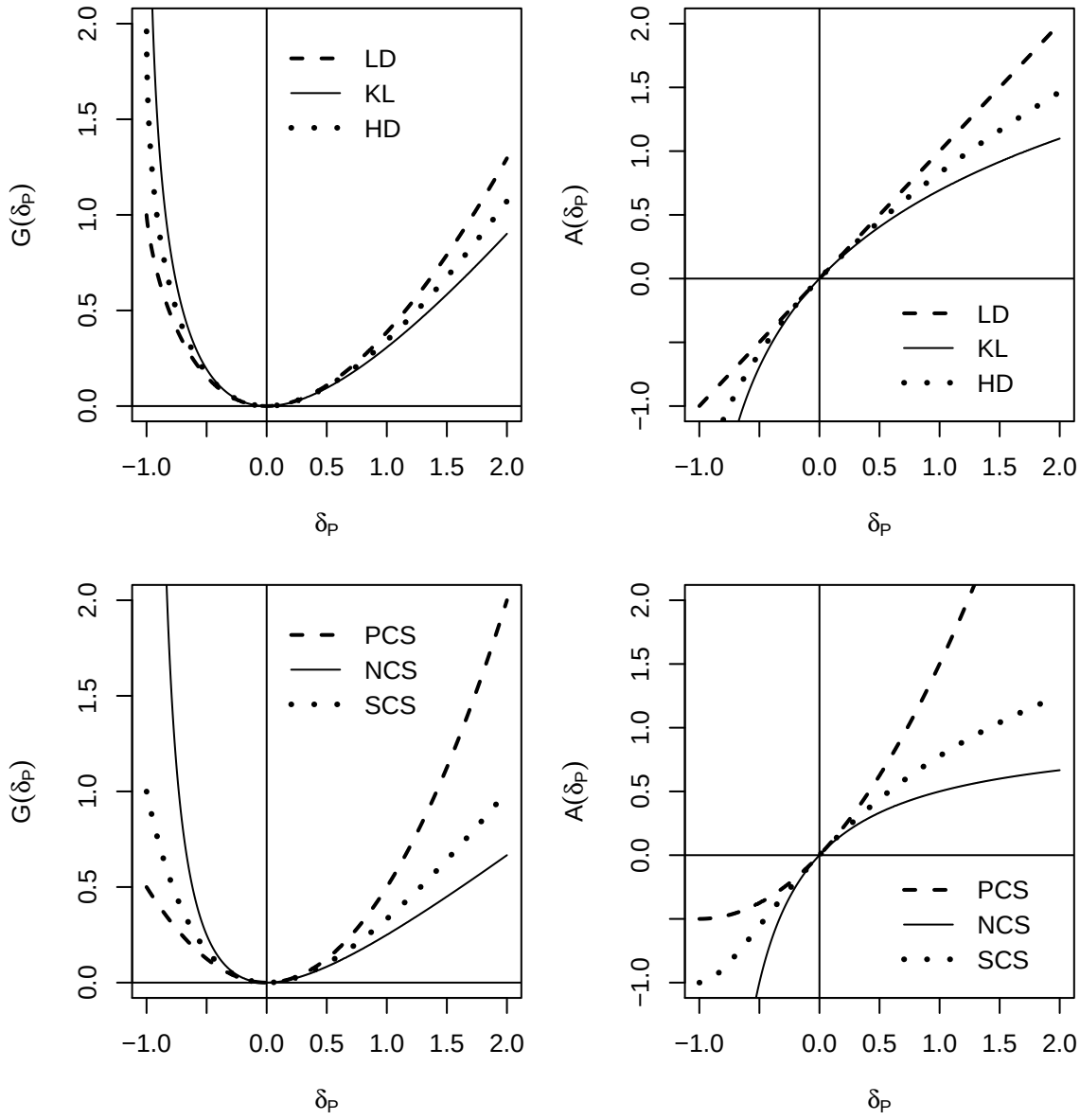

Figure 2.2: G and A functions for popular distances

One way to compare two disparities is to study the ratio between residual adjustment functions. Suppose that $A_1(\cdot)$ and $A_2(\cdot)$ are the RAF's for disparities ρ_1 and ρ_2, respectively. Then the quantity of $A_1(\delta_P)/A_2(\delta_P)$ will give insight on the similarities and differences between ρ_1 and ρ_2. For example, when $A_1(\delta_P)/A_2(\delta_P) < 1$

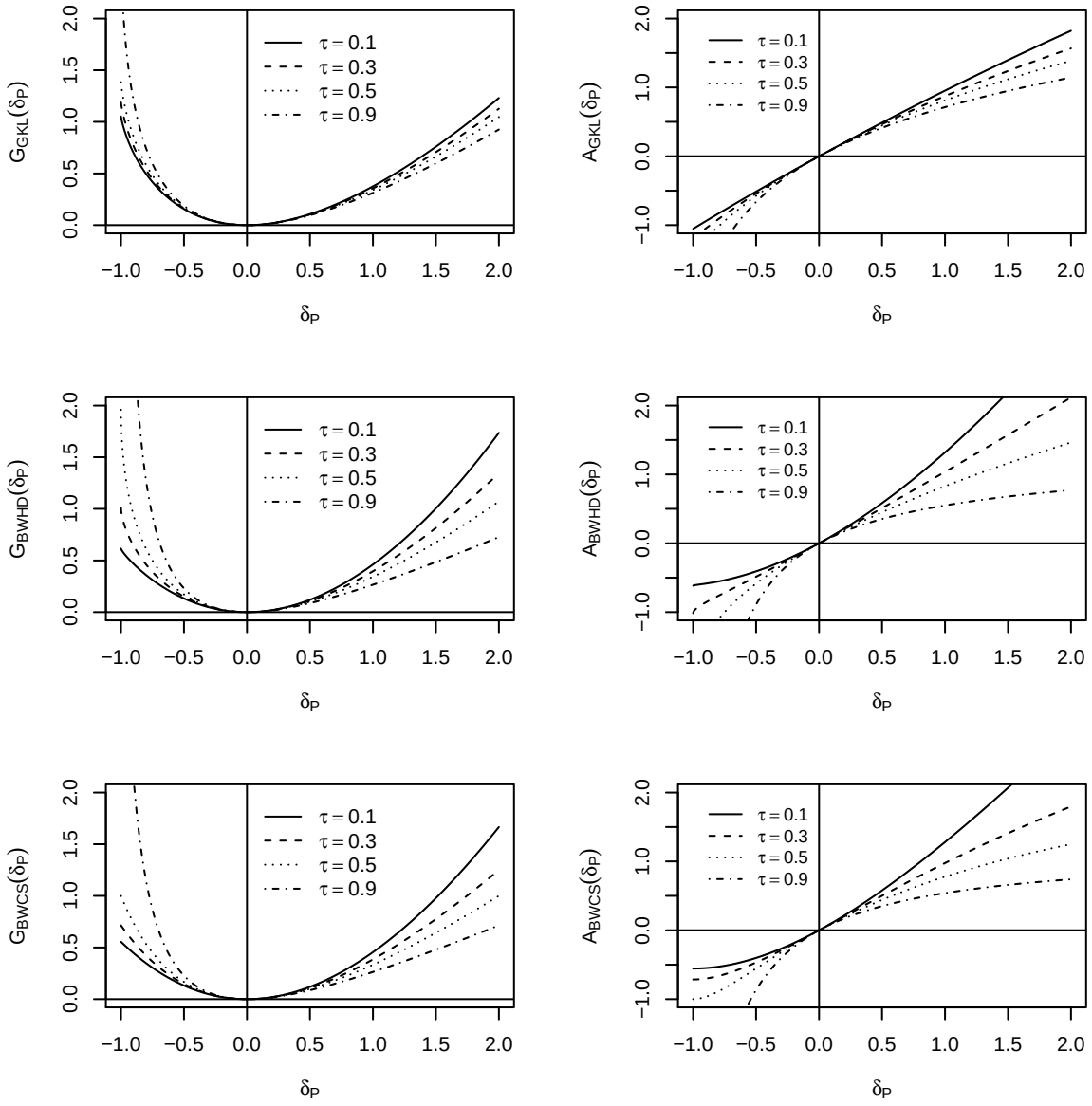

Figure 2.3: G and A functions for popular generalized distances

the weight is greater for ρ_2 at a certain value of the Pearson residual δ_P. Similarly, when $A_1(\delta_P)/A_2(\delta_P) > 1$, the weight is greater for ρ_1 at a certain value of the Pearson residual δ_P. Since the maximum likelihood estimator is typically the status quo in statistical inference, it makes sense to compare disparities with the RAF corresponding

to the likelihood disparity (LD).

The weighted likelihood function for a disparity with RAF function $A(\cdot)$ is defined in [Lindsay, 1994] as

$$w(\delta_P(y)) = \frac{A(\delta_P(y))}{A_{LD}(\delta_P(y))} = \frac{A(\delta_P(y))}{\delta_P(y)}$$

and its purpose is to compare the weights applied to residuals in a minimum disparity procedure with the maximum likelihood procedure. By design, distances with $\lim_{\delta_P \to -1^+} w(\delta_P) = 0$ are inlier robust and distances with $\lim_{\delta_P \to \infty} w(\delta_P) = 0$ are robust. Unfortunately, this particular weighted likelihood function is still defined on an unbounded interval and, as a result, the limits for the weight function can be difficult to compute and the graphs may be misleading.

In [Park et al., 2002], an alternative approach was introduced which defined the weighted likelihood function on the interval $[-1, 1]$. Let Neyman's residual be defined as

$$\delta_N(x) = \frac{d(x) - f_\theta(x)}{d(x)}$$

and the combined residual is defined in [Park et al., 2002] as

$$\delta_C(x) = \begin{cases} \delta_P(x) & : \text{if} \quad d(x) \leq f_\theta(x) \\ \delta_N(x) & : \text{if} \quad d(x) \geq f_\theta(x) \end{cases}.$$

where $\delta_C \in [-1, 1)$. The weighted likelihood function redefined on δ_C then becomes

$$w_C(\delta_C) = \begin{cases} \frac{A(\delta_C)}{\delta_C} & : \text{if} \quad -1 \leq \delta_C < 0, \\ A'(0) & : \text{if} \quad \delta_C = 0 \\ \frac{1-\delta_C}{\delta_C} A(\frac{\delta_C}{1-\delta_C}) & : \text{if} \quad 0 < \delta_C < 1, \\ A'(\infty) & : \text{when} \quad \delta_C = 1 \end{cases} .$$

A simplified version of this weight function will be introduced later in this section. The weighted likelihood functions for the popular distances defined earlier are shown in Figure 2.4.

Notice that by design the weight function for the likelihood disparity is just $w_{LD}(\delta_C) = 1$. When looking at the weighted likelihood functions shown in Figure 2.4, we can use the likelihood disparity as a benchmark to compare with the other disparities. So for a particular disparity, we only need to check the behavior at $\delta_C = -1$ and $\delta_C = 1$. That is, $w(-1) = 0$ implies inlier robustness and $w(1) = 0$ implies robustness. We can see that both KL and HD are robust, but not inlier robust. The PCS is not robust and the NCS is not inlier robust. GKL, BWHD, and BWCS all become more robust (and less anti-robust) as τ increases.

Although the issues concerning the unbounded domain were fixed, it still uses Neyman's residual on one part of the interval and Pearson's residual on the other part. To obtain a weight function defined on a single bounded domain, we offer the following suggestion. Let

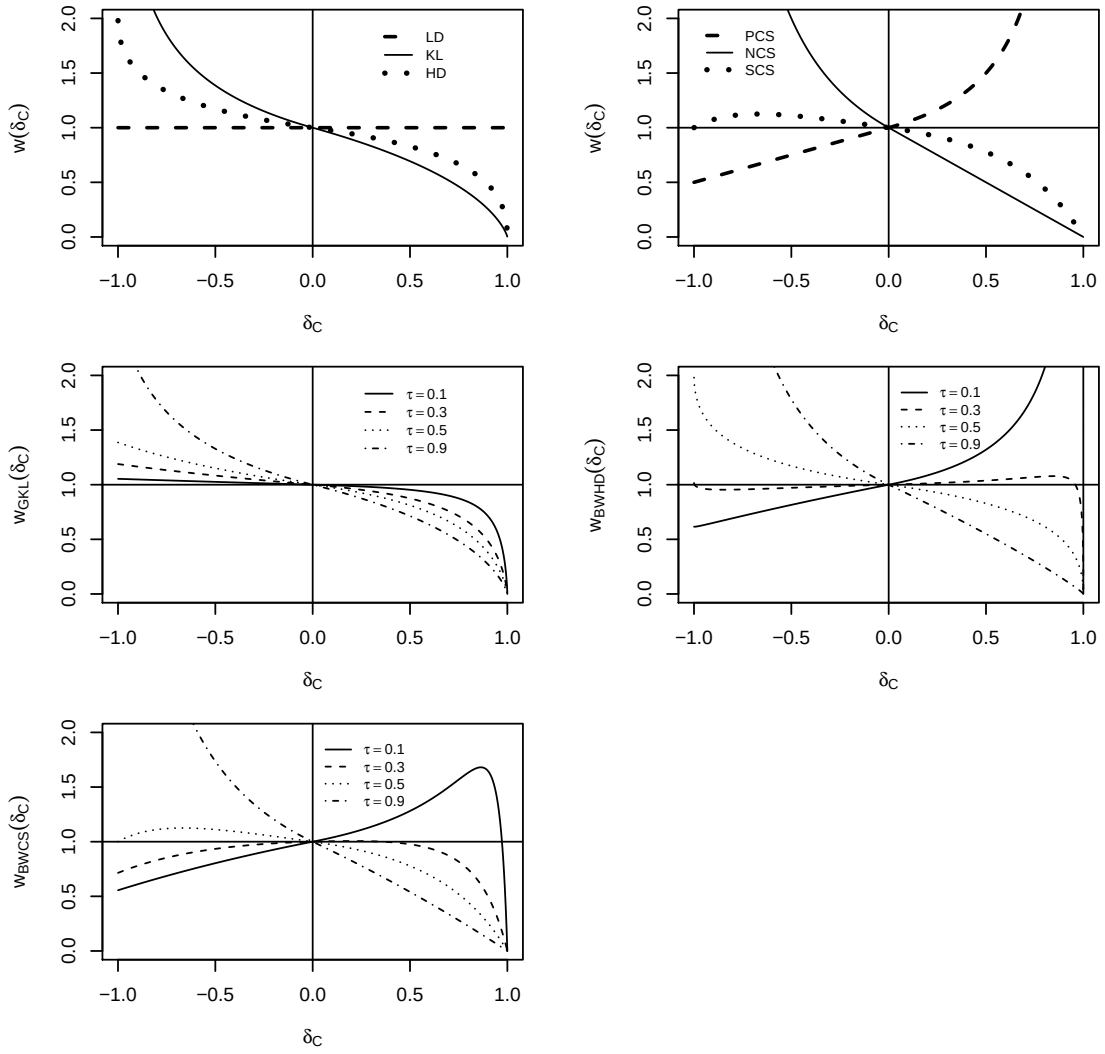

Figure 2.4: Weighted likelihood functions for popular statistical distances

$$\delta_S = \frac{d(y) - f_\theta(y)}{d(y) + f_\theta(y)} = \frac{\delta_P(y)}{\delta_P(y) + 2}$$

represent the symmetric residual. Note that since $\delta_P \in [-1, \infty)$ it means that $\delta_S \in [-1, 1)$. The new weighted likelihood function can now be defined as

$$
\begin{aligned}
w(\delta_S) &= \frac{A(\delta_P(\delta_S))}{\delta_P(\delta_S)} \\
&= \frac{A\left(\frac{2\delta_S}{1-\delta_S}\right)}{\frac{2\delta_S}{1-\delta_S}} \\
&= \frac{(1-\delta_S)A\left(\frac{2\delta_S}{1-\delta_S}\right)}{2\delta_S}.
\end{aligned}
$$

Furthermore, in our future work we will show that the endpoints for the weighted likelihood function defined on this symmetric residual will be the same as the weighted likelihood function defined on the combined residual. The plots for this function of popular disparities are shown in Figure 2.5. It is clear that Figures 2.4 and 2.5 are close to identical. In particular, the new weight function gives the same three-point summary which can be used to determine the robustness and efficiency of a disparity measure.

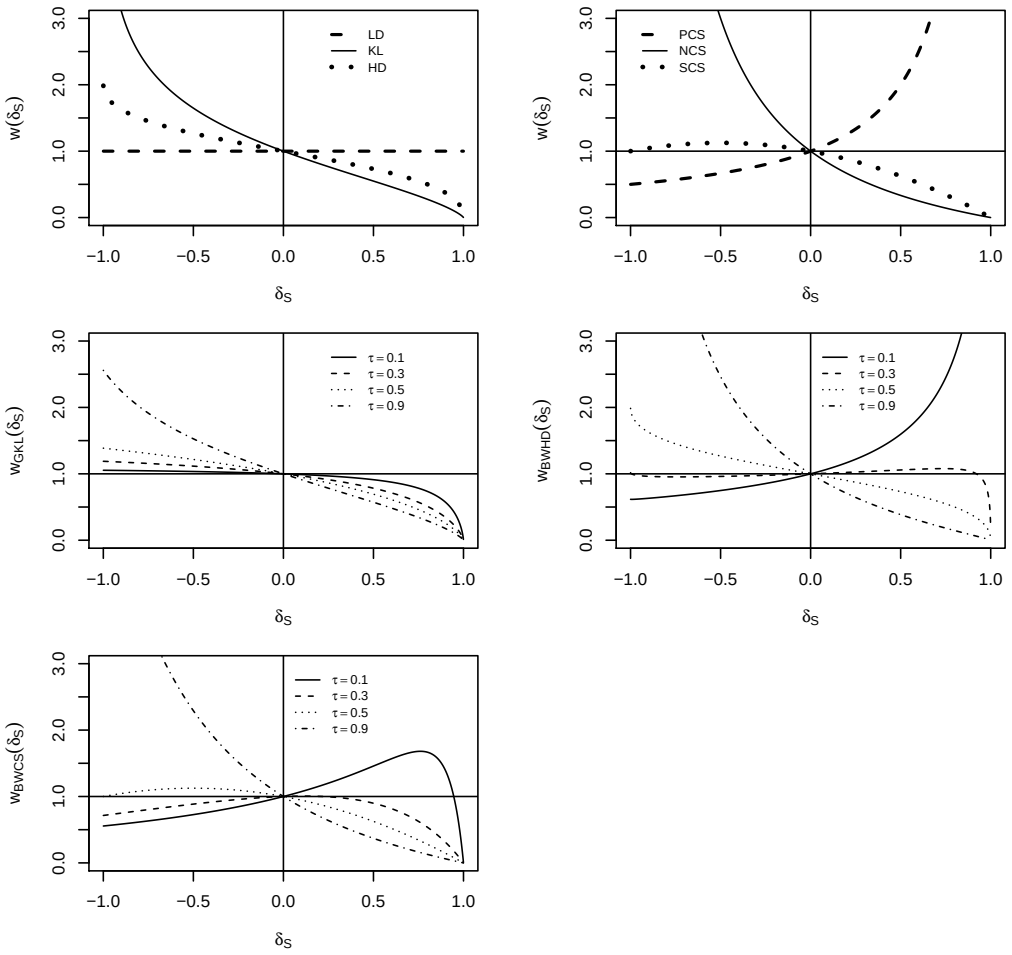

Figure 2.5: Weighted likelihood functions for popular statistical distances based on symmetric residual

Chapter 3

Symmetrization of Disparities

3.1 New Disparity-based Distances

In this chapter, we will introduce a new class of statistical distances that can be used to analyze a wide range of problems. The distances mentioned in this chapeter are just a few of many distances that can be used for statistical inference. So far we have looked at the RAF and the weighted likelihood function, which both tell us whether or not the disparity estimator will have desirable properties (efficiency and robustness). Unfortunately, the applications that we will be looking at do not involve parameter estimation. Instead we wish to quantify the discrepancy between two statistical objects and we need our distance to be symmetric.

At this point, we have only seen a few symmetric distances: Hellinger (HD), symmetric chi-square (SCS), and GKL with $\tau = 0.5$. Now the goal is to obtain generalized versions of these symmetric distances. The idea is simple, for any disparity measure ρ the symmetric version will be defined as

$$\rho^* = \frac{\rho(d, f_\theta) + \rho(f_\theta, d)}{2} \tag{3.1}$$

By design, the new distance ρ^* will also be a disparity and therefore can be used in goodness of fit testing. Since the disparity ρ^* is also symmetric, it can be used in a wider range of applications. There are many disparity measures to choose from but this paper focuses on symmetric versions of the GKL, BWHD, and BWCS distances introduced in the previous chapter. The first class of symmetric distances to be considered is the symmetric generalized Kullback-Leibler, and will be defined as

$$\text{SGKL}_\tau(d, f_\theta) = \frac{\text{GKL}_\tau(d, f_\theta) + \text{GKL}_\tau(f_\theta, d)}{2}$$

Note that $\text{SGKL}_{\tau=0.5} = \text{GKL}_{\tau=0.5}$. The next class of symmetric distances to be considered is the symmetric blended-weight Hellinger distance, and will be defined as

$$\text{SBWHD}_\tau(d, f_\theta) = \frac{\text{BWHD}_\tau(d, f_\theta) + \text{BWHD}_\tau(f_\theta, d)}{2}$$

Note that $\text{SBWHD}_{\tau=0.5} = \text{BWHD}_{\tau=0.5} = HD$. The last class of symmetric distances to be considered is the symmetric blended-weight chi-square distance, and will be defined as

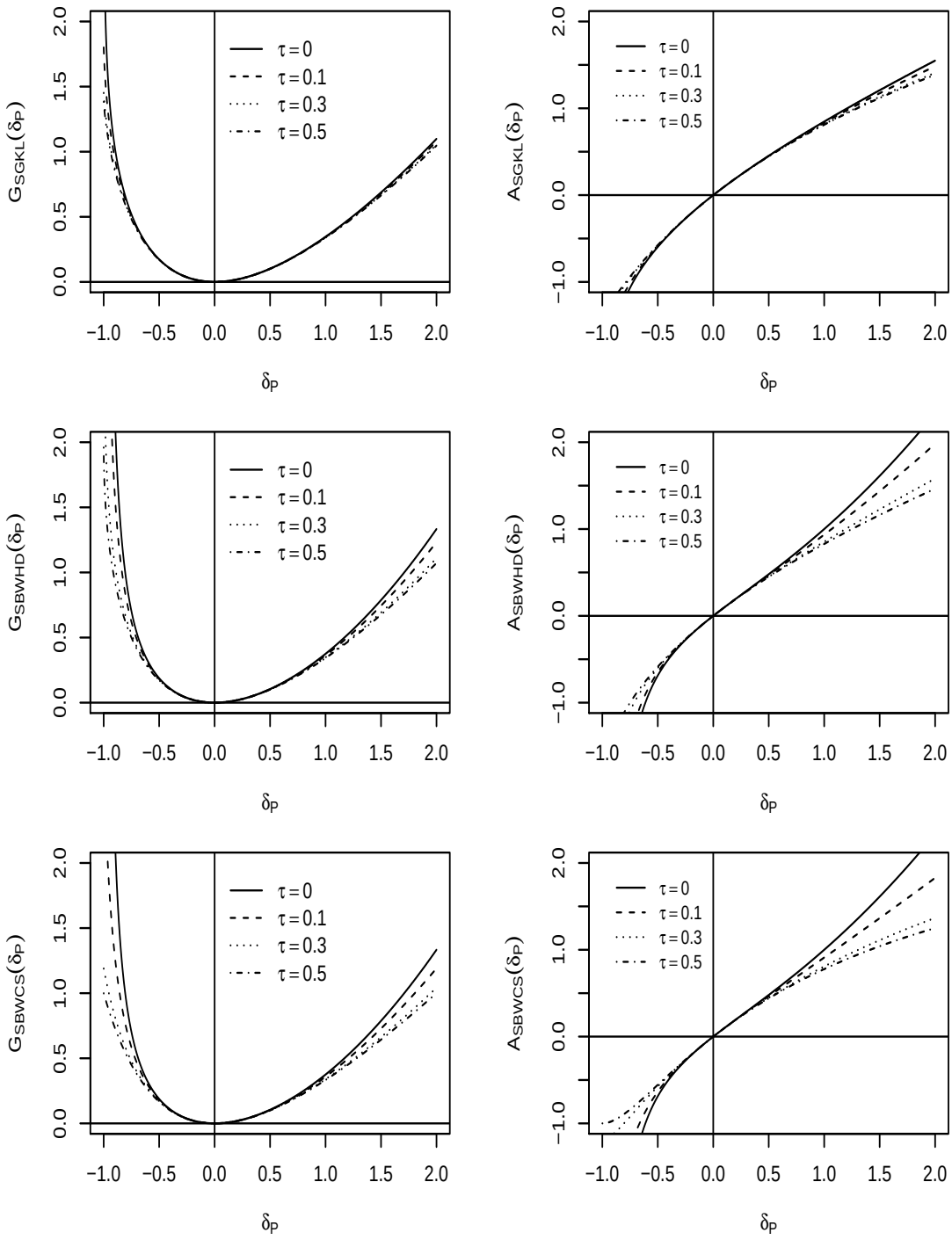

Figure 3.1: G and A functions for symmetric disparities

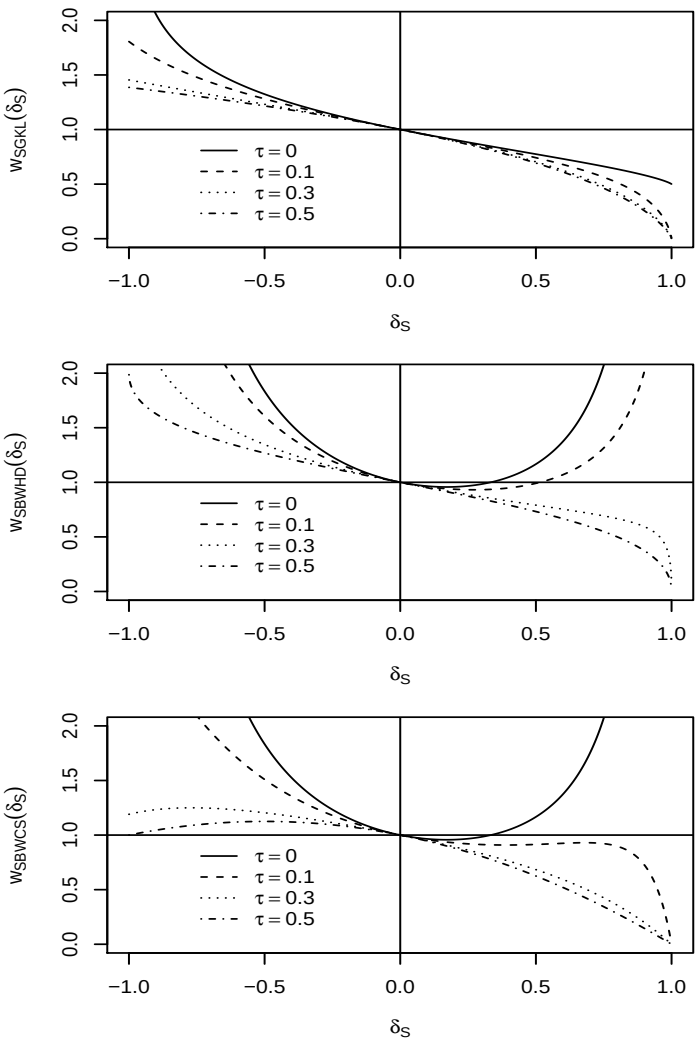

Figure 3.2: w functions for symmetric disparities

$$\text{SBWCS}_\tau(d, f_\theta) = \frac{\text{BWCS}_\tau(d, f_\theta) + \text{BWCS}_\tau(f_\theta, d)}{2}$$

Note that $\text{SBWCS}_{\tau=0.5} = \text{BWCS}_{\tau=0.5} = SCS$. The G and RAF functions for these symmetric distances are shown in Figure 3.1 and the weighted likelihood functions are shown in Figure 3.2. The properties of these distances can be analyzed following the same three-point summary defined in the previous chapter. Figures 3.1 and 3.2 show that none of the distances are inlier robust since $w(-1) > 0$. Also Figures 3.1 and 3.2 show that SGKL_τ $(0 \leq \tau \leq 1)$, SBWHD_τ $(\tau > 0.1)$, and SBWCS_τ $(\tau > 0)$ are robust since $w(+1) = 0$.

3.2 New G Function

Consider representing the disparity measure ρ using the symmetric residual

$$\delta_S(x) = \frac{d(x) - f_\theta(x)}{d(x) + f_\theta(x)} = \frac{\delta_P(x)}{\delta_P(x) + 2}$$

Now instead of using Pearson's residual, we will try rewriting our disparities using the symmetric residual:

$$\begin{aligned}
\rho_G(d, f_\theta) &= \sum_{i=1}^{k} G(\delta_P(y_i)) f_\theta(y_i) \\
&= \sum_{i=1}^{k} G(\delta_P(y_i)) \frac{f_\theta(y_i)}{f_\theta(y_i) + d(y_i)} (f_\theta(y_i) + d(y_i)) \\
&= \sum_{i=1}^{k} G(\delta_P(y_i)) \Big(\frac{1}{\delta_P + 2}\Big) (f_\theta(y_i) + d(y_i)) \\
&= \sum_{i=1}^{k} G^*(\delta_P(y_i)) (f_\theta(y_i) + d(y_i))
\end{aligned}$$

where we let

$$G^{\rho}_{sym}(\delta_S) = G^*(\delta_P) = \frac{G(\delta_P)}{\delta_P + 2} = \frac{G\Big(\frac{2\delta_S}{1-\delta_S}\Big)}{\frac{2\delta_S}{1-\delta_S} + 2} = \frac{1}{2} G\Big(\frac{2\delta_S}{1-\delta_S}\Big)(1 - \delta_S)$$

Consider the form of $G^{\rho}_{sym}(\delta_S)$ for some of the popular disparities defined in Chapter 2. A summary of $G^{\rho}_{sym}(\delta_S)$ functions is shown in Table 3.1. Figure 3.3 shows that the new $G^{\rho}_{sym}(\delta_S)$ function is symmetric for the symmetric versions of the chi-square, Hellinger, and generalized Kullback-Leibler distances.

In parameter estimation, the likelihood and Pearson's chi-square distances are very popular and often used as a reference to compare other distances. An example of this was introduced in the previous chapter when a weight function, using the likelihood distance as a reference, was examined. It is now necessary to consider new references since our focus is on applications that require symmetry. Consider a weighted version of the G function where we consider the symmetric Chi-square, Hellinger, and generalized Kullback-Leibler distances.

Disparity	$G^*_\rho(\delta_S)$	Symmetric?
LD	$\frac{1}{2}(1+\delta_S)\log\left(\frac{1+\delta_S}{1-\delta_S}\right)-\delta_S$	No
KL	$-\frac{1}{2}(1-\delta_S)\log\left(\frac{1+\delta_S}{1-\delta_S}\right)+\delta_S$	No
PCS	$\frac{\delta_S^2}{1-\delta_S}$	No
NCS	$\frac{\delta_S^2}{1+\delta_S}$	No
SCS	δ_S^2	Yes
HD	$2-2\sqrt{(1+\delta_S)(1-\delta_S)}$	Yes
$\text{GKL}_{\tau=0.5}$	$(1+\delta_S)\log\left(\frac{1+\delta_S}{1-\delta_S}\right)-2\log\left(\frac{1}{1-\delta_S}\right)$	Yes

Table 3.1: G^* functions for popular disparities

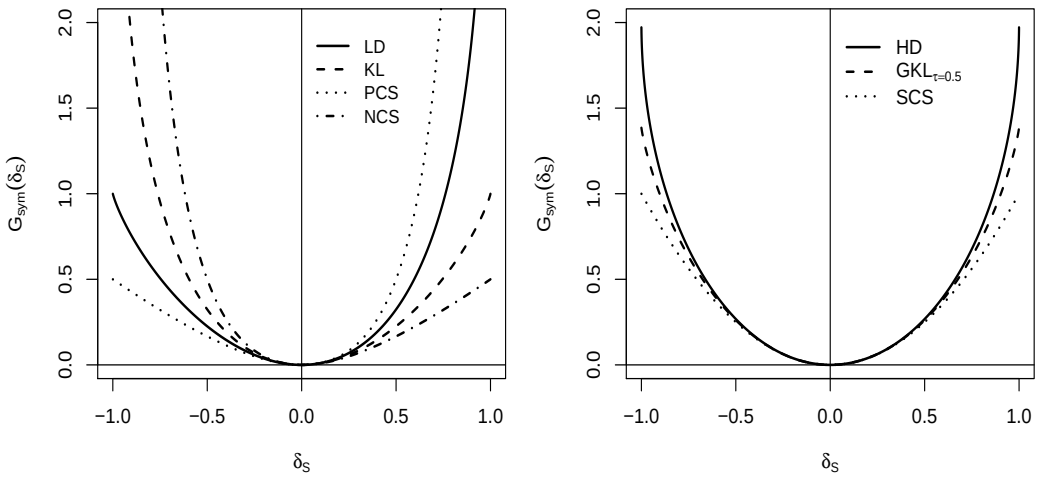

Figure 3.3: G^* functions for popular disparities

For example, we can use either of the following weight functions for investing the properties of a general distance ρ:

$$w^*_{\text{SCS}}(\delta_S) = \frac{G^{\rho}_{sym}(\delta_S)}{G^{\text{SCS}}_{sym}(\delta_S)}$$

$$w^*_{\text{HD}}(\delta_S) = \frac{G^{\rho}_{sym}(\delta_S)}{G^{\text{HD}}_{sym}(\delta_S)}$$

$$w^*_{\text{GKL}}(\delta_S) = \frac{G^{\rho}_{sym}(\delta_S)}{G^{\text{GKL}}_{sym}(\delta_S)}$$

The weighted G functions are shown in Figure 3.4. It is clear that for symmetric distances, Hellinger (HD), symmetric chi-square (SCS), and GKL with $\tau = 0.5$, the weight functions are also symmetric for each weight. Figures 3.5, 3.6, and 3.7 clearly show that the new symmetric distances will have symmetric $G^{\rho}_{sym}(\delta_S)$ and $w^*(\delta_S)$ for any value of τ.

Depending on the reference function used in the weight function w^*, it is natural to think that conclusions can be drawn based on the curvature about whether or not there is down-weighting for large discrepancies. This will be investigated in future work, but still we are still encouraged by the fact that this new weight function does emphasize symmetry.

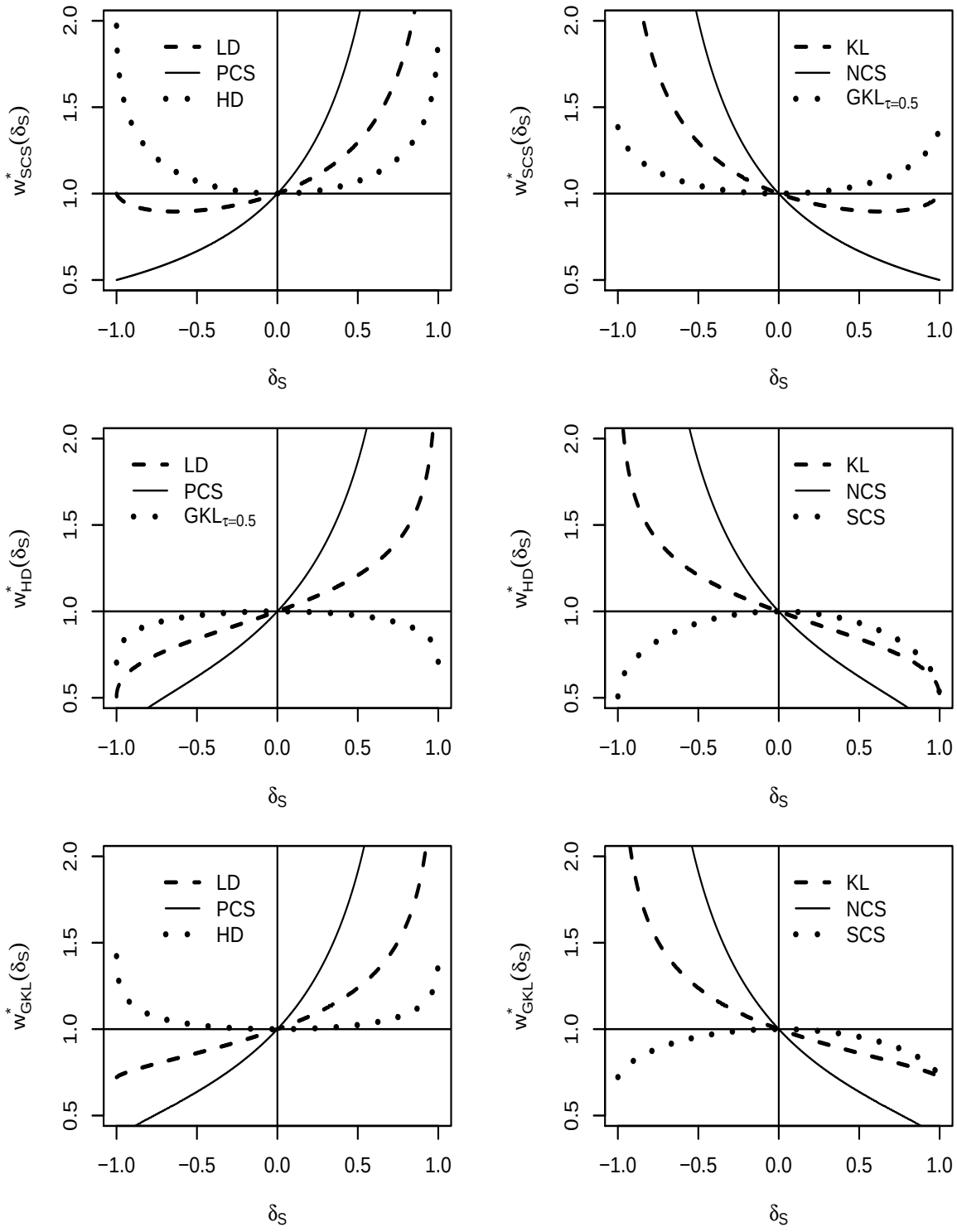

Figure 3.4: New w^* functions for popular disparities weighted by SCS, HD, and $\text{GKL}_{\tau=0.5}$

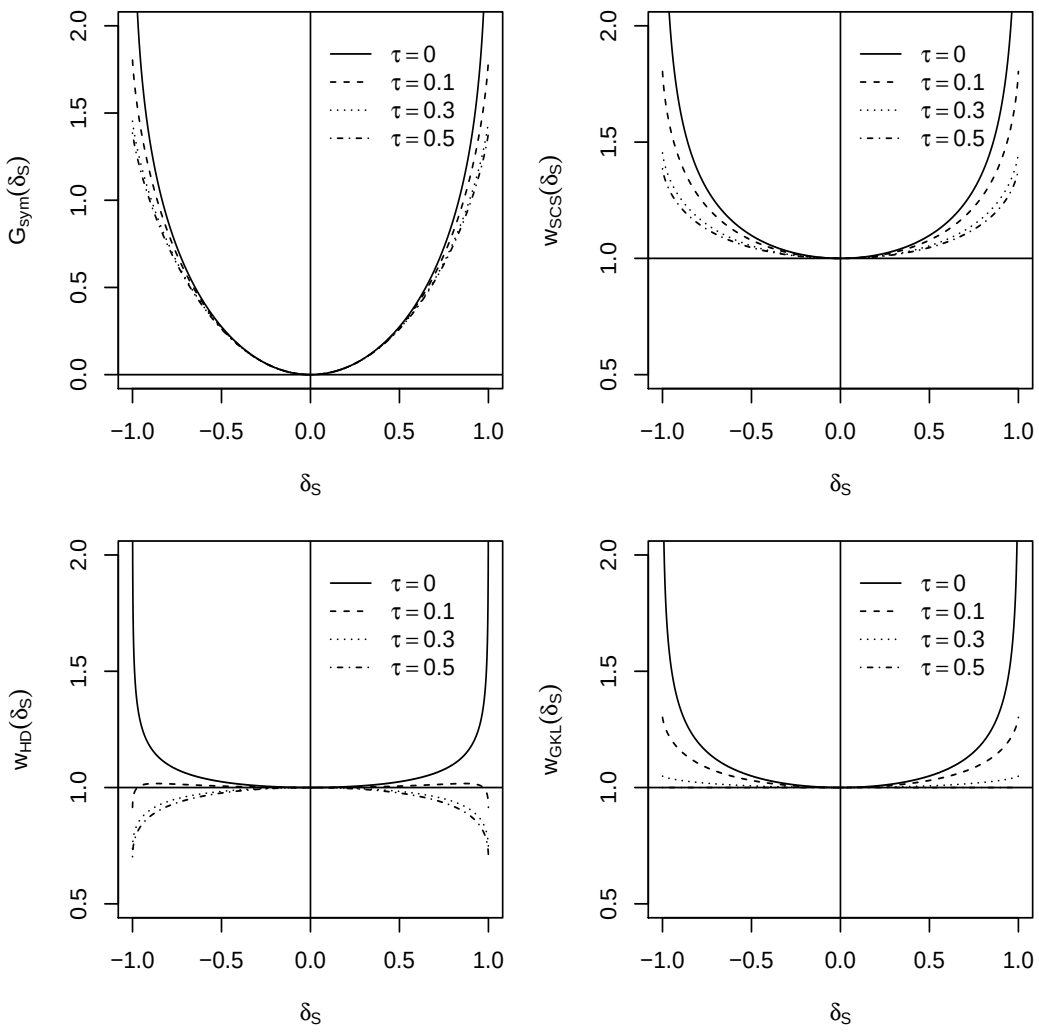

Figure 3.5: SGKL

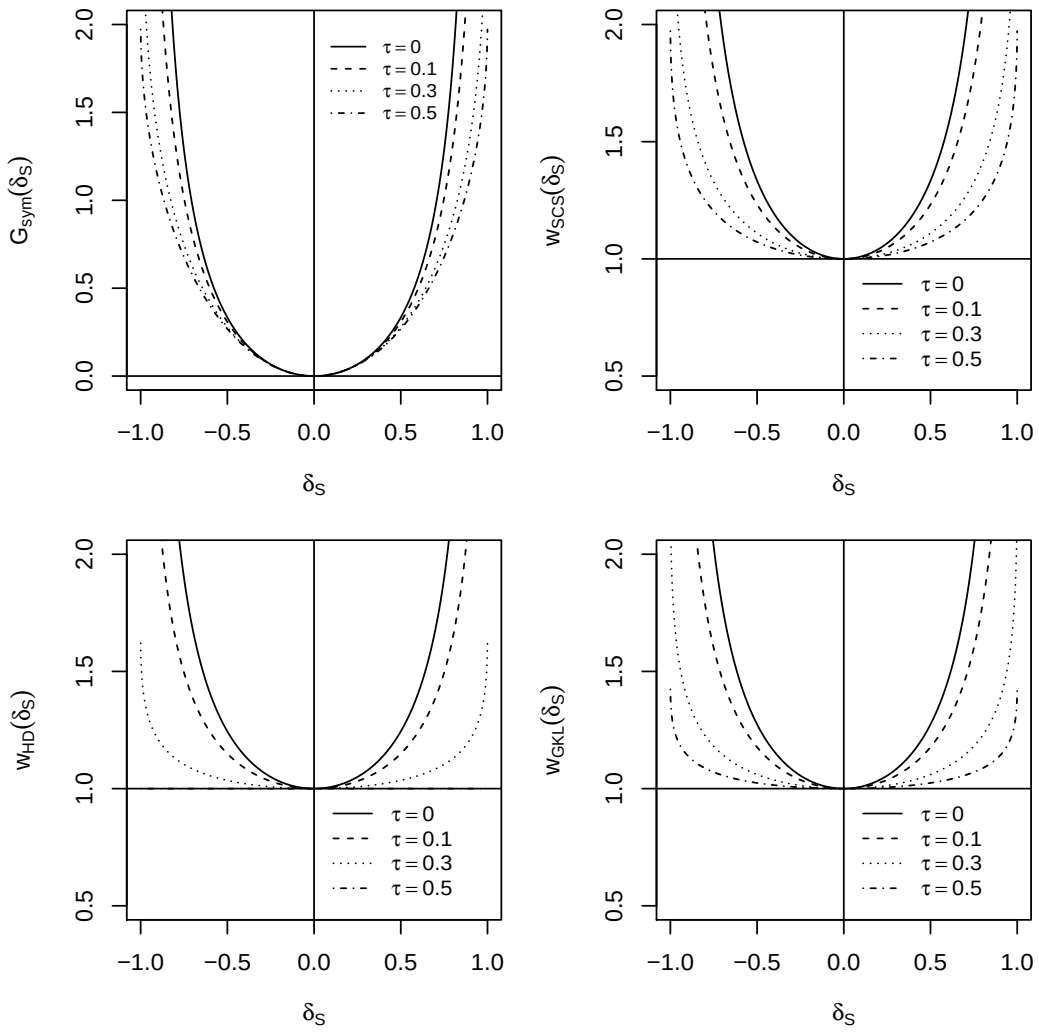

Figure 3.6: SBWHD

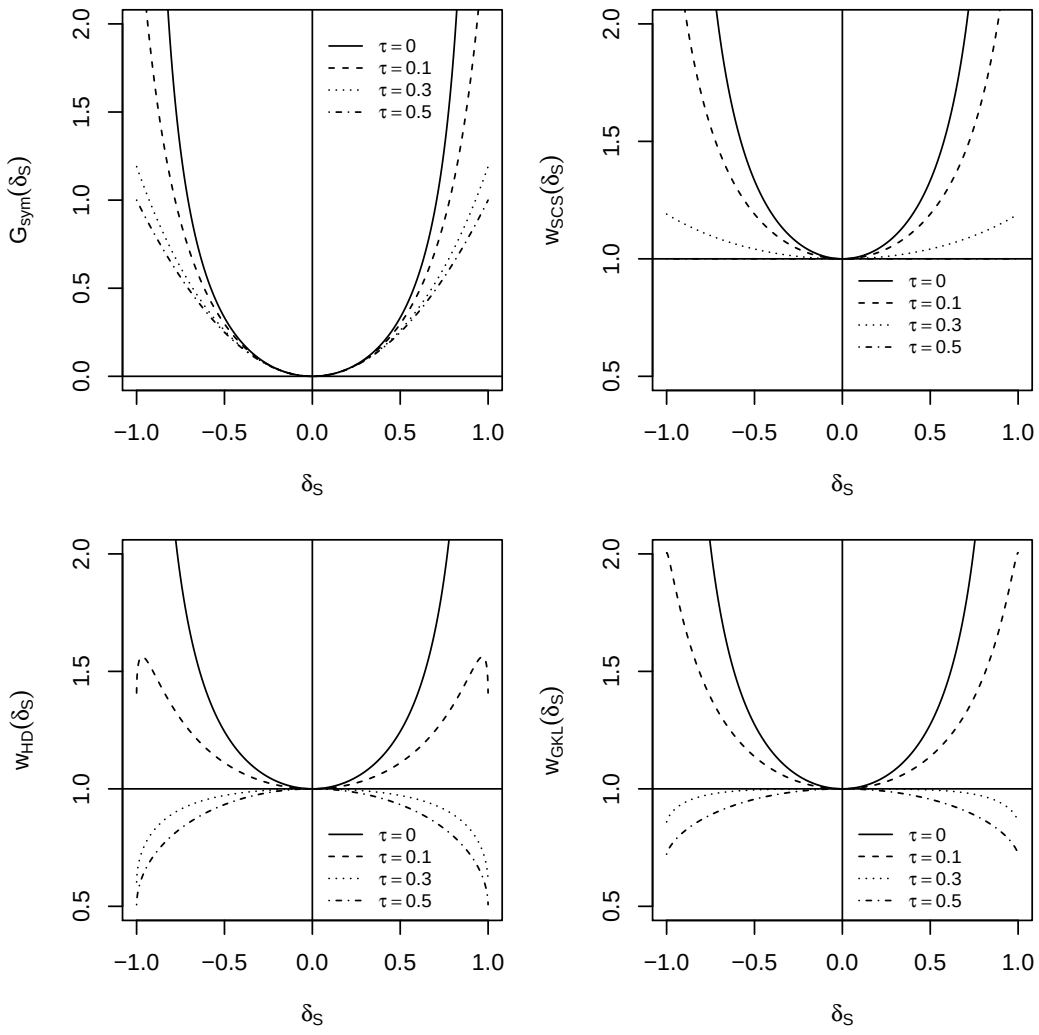

Figure 3.7: SBWCS

Chapter 4

Ecology Problem

4.1 Species Composition Data

The collection of all living things within a specific ecosystem or area is referred to as the species composition. In ecology, it is important to understand the variation in species composition within a particular site and among other sites within a region of interest. In [Whittaker, 1972], these variations became known as alpha, beta, and gamma diversities. Alpha diversity is the variation of species at a site, beta diversity is the variation of species among sites within a region, and gamma diversity is the variation of species within an entire region [Whittaker, 1972]. At the moment, our focus will be on the beta diversity. This variation can be analyzed by calculating an ecological distance between two sites. When comparing two sites that share many of the same species, the ecological distance between the sites should be small. Otherwise, if the sites have few species in common, the ecological distance should be large. Once we have chosen an appropriate ecological distance, we calculate this distance between all pairs of sites in the study and record these values in what we call a distance or dissimilarity matrix. Figure 4.1 shows the typical approach to the ecology problem

we are interested in.

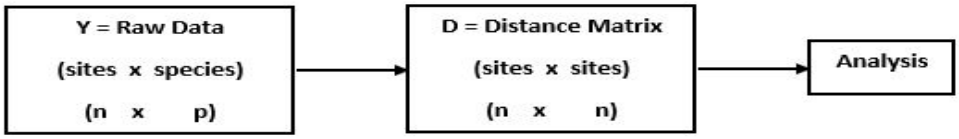

Figure 4.1: Ecology Problem

There are many methods for analyzing a species composition data set. This paper will focus on comparing the functions used to obtain the distance matrix, and then look at different methods of analysis.

4.2 Distances in Ecology

Let D represent the distance matrix and $D(a, b)$ represent the distance between sites a and b based on species composition. According to [Legendre and Legendre, 2012], the following properties are necessary for all distances in ecology in measuring beta diversity:

- Positiveness: $D \geq 0$
- Symmetry: $D(a, b) = D(b, a)$
- Monotonicity: D increases as the differences in abundance increases.
- Double Zero Asymmetry: D does not change when adding double zeros, but D changes when sites have a species in common.
- Sites without species in common have the largest D.
- D does not decrease in series of nested species collection.

As you can imagine, there are many possible distances for us to consider when comparing sites. We will start by introducing some of the most popular distances in ecology, including the Euclidean distance, chi-square based distances, and Hellinger based distances. When analyzing the differences in species composition, many of the classical approaches can be poor choices for certain data sets. For example, consider the following species abundance data from [Legendre and Gallagher, 2001]

Table 4.1: Species Abundance Paradox

	Species 1	Species 2	Species 3
Site 1	0	1	1
Site 2	1	0	0
Site 3	0	4	8

It is natural to consider a Euclidean distance to measure the distance between two ecological sites. In [Rao, 1995], the Hellinger distance was analyzed using species abundance data. The Euclidean and Hellinger based distance matrices are shown below in Table 4.2. Later we will define each ecological distance, but for now we can compare the matrices for this example.

Table 4.2: Euclidean and Hellinger Distance Matrices

Euclidean			
	Site 1	Site 2	Site 3
Site 1	0.0000	1.7321	7.6158
Site 2	1.7321	0.0000	9.0000
Site 3	7.6158	9.0000	0.0000

Hellinger			
	Site 1	Site 2	Site 3
Site 1	0.0000	1.4142	0.1697
Site 2	1.4142	0.0000	1.4142
Site 3	0.1697	1.4142	0.0000

It is easy to see from Table 4.1 that sites 1 and 3 have the most in common since they share the same species. However, the Euclidean distance suggests that the largest discrepancy is between sites 2 and 3. This is because the Euclidean distance is highly effected by the species abundances. In contrast, when looking at the Hellinger distance matrix we calculate a small discrepancy between sites 1 and 3, and the site pairs with no species in common both receive a maximum distance of $\sqrt{2} \approx 1.4142$. This example highlights the fact that looking at only species abundances can be a little misleading, and of greater importance is whether or not two sites share the same relative frequencies. Also it is clear from this example that the classical Euclidean distance is not appropriate for species abundance data. Therefore, it would be beneficial to explore more distances that can be used in this area to improve our understanding of the problem.

Consider a species abundance data set with n sites and p species. Let a_j and b_j represent the number of species $j = 1, \ldots, p$ at sites a and b, respectively. Also let $a_+ = \sum_{j=1} a_j$ and $b_+ = \sum_{j=1} b_j$ represent the total number of species at sites a and b, respectively. Then the following distances are some of the most popular for finding beta diversity [Legendre and Gallagher, 2001]

$$D_{Euclidean}(a,b) = \sqrt{\sum_{j=1}^{p} \left(a_j - b_j\right)^2}$$

$$D_{Hellinger}(a,b) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{a_j}{a_+}} - \sqrt{\frac{b_j}{b_+}}\right]^2}$$

$$D_{chord}(a,b) = \sqrt{\sum_{j=1}^{p}\left(\frac{a_j}{\sqrt{\sum_{j=1}^{p} a_j^2}} - \frac{b_j}{\sqrt{\sum_{j=1}^{p} b_j^2}}\right)^2}$$

$$D_{\chi^2 distance}(a,b) = \sqrt{\sum_{j=1}^{p} \frac{a_+ + b_+}{a_j + b_j}\left(\frac{a_j}{a_+} - \frac{b_j}{b_+}\right)^2}$$

$$D_{speciesprofiles}(a,b) = \sqrt{\sum_{j=1}^{p}\left(\frac{a_j}{a_+} - \frac{b_j}{b_+}\right)^2}$$

$$D_{BrayCurtis}(a,b) = 1 - 2\frac{\sum_{j=1}^{p} min(a_j, b_j)}{a_+ + b_+}$$

$$D_{Kulczynski}(a,b) = 1 - \frac{1}{2}\left(\frac{\sum_{j=1}^{p} min(a_j, b_j)}{a_+} + \frac{\sum_{j=1}^{p} min(a_j, b_j)}{b_+}\right)$$

Shown below are the ecology versions of the symmetric distances defined in the previous chapter:

$$\mathrm{SGKL}_\tau(a,b) = \sqrt{\frac{1}{2}\Big[\mathrm{GKL}_\tau\Big(\frac{a}{a_+}, \frac{b}{b_+}\Big) + \mathrm{GKL}_\tau\Big(\frac{b}{b_+}, \frac{a}{a_+}\Big)\Big]}$$

$$\mathrm{SBWHD}_\tau(a,b) = \sqrt{\frac{1}{2}\Big[\mathrm{BWHD}_\tau\Big(\frac{a}{a_+}, \frac{b}{b_+}\Big) + \mathrm{BWHD}_\tau\Big(\frac{b}{b_+}, \frac{a}{a_+}\Big)\Big]}$$

$$\mathrm{SBWCS}_\tau(a,b) = \sqrt{\frac{1}{2}\Big[\mathrm{BWCS}_\tau\Big(\frac{a}{a_+},\frac{b}{b_+}\Big) + \mathrm{BWCS}_\tau\Big(\frac{b}{b_+},\frac{a}{a_+}\Big)\Big]}$$

In the next section a comparison will be made between the new class of distances and the classical approaches in ecology. It is our belief that the methods proposed in this paper will be very useful in comparing species composition due because of flexibility. Also during our comparison of difference ecological distances, we believe this paper will offer some useful insight as to how one could choose which distance to use for a particular set of species composition data.

4.3 Analysis of Artificial Gradient Data

In order to compare distances in ecology, it is necessary to know the actual geographic data between sites and then compare with the distance matrix. This can be done using what we call artificial gradient data [Legendre and Gallagher, 2001]. The idea is we arrange the n sites in the sequence of $1-n$ based on the similiarities between sites. Specifically, the true geographic (or gradient) distance between sites a and b is just $|a-b|$.

Consider the following species composition data from [Kindt and Coe, 2005] shown in Table 4.3 and Figure 4.2:

One could start the comparison by looking at Hellinger's distance, the corresponding distance matrix for the artificial gradient data is shown in Table 4.4.

A useful graphical representation of the distance matrix is plotting the Hellinger distances versus the true geographic distance [Legendre and Gallagher, 2001]. This type representation allows us to easily see whether or not a distance is monotone.

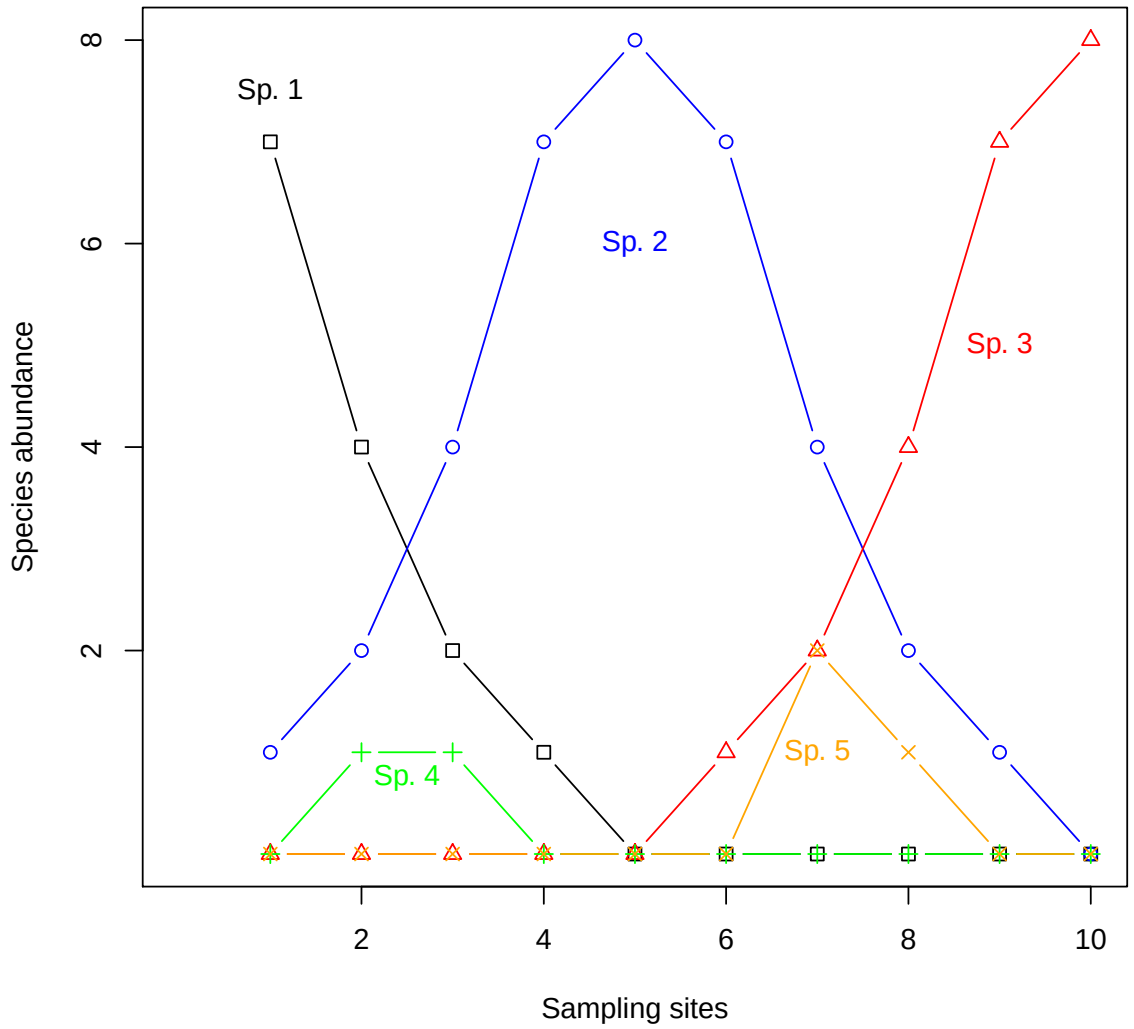

Figure 4.2: Artificial dataset with the abundance of 5 species for 10 sites

For each true distance, we will use a single point index to compare will be the average distance calculated by the chosen function. Let $f(i)$ represent the diastegram function when the true geographic distance is i, then by definition

Table 4.3: Artificial dataset with the abundance of 5 species for 10 sites

	Species 1	Species 2	Species 3	Species 4	Species 5
Site 1	7	1	0	0	0
Site 2	4	2	0	1	0
Site 3	2	4	0	1	0
Site 4	1	7	0	0	0
Site 5	0	8	0	0	0
Site 6	0	7	1	0	0
Site 7	0	4	2	0	2
Site 8	0	2	4	0	1
Site 9	0	1	7	0	0
Site 10	0	0	8	0	0

Table 4.4: Hellinger Distance Matrix for Artificial Gradient Data

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Site 9	Site 10
Site 1	0.0000	0.4559	0.6823	0.8229	1.1371	1.1570	1.2247	1.2736	1.3229	1.4142
Site 2	0.4559	0.0000	0.3131	0.6823	0.9649	1.0000	1.1154	1.1952	1.2736	1.4142
Site 3	0.6823	0.3131	0.0000	0.4559	0.6987	0.7654	0.9649	1.0917	1.2106	1.4142
Site 4	0.8229	0.6823	0.4559	0.0000	0.3594	0.5000	0.8229	1.0000	1.1570	1.4142
Site 5	1.1371	0.9649	0.6987	0.3594	0.0000	0.3594	0.7654	0.9649	1.1371	1.4142
Site 6	1.1570	1.0000	0.7654	0.5000	0.3594	0.0000	0.5688	0.6823	0.8229	1.1371
Site 7	1.2247	1.1154	0.9649	0.8229	0.7654	0.5688	0.0000	0.3319	0.7514	1.0000
Site 8	1.2736	1.1952	1.0917	1.0000	0.9649	0.6823	0.3319	0.0000	0.4559	0.6987
Site 9	1.3229	1.2736	1.2106	1.1570	1.1371	0.8229	0.7514	0.4559	0.0000	0.3594
Site 10	1.4142	1.4142	1.4142	1.4142	1.4142	1.1371	1.0000	0.6987	0.3594	0.0000

$$f(i) = \frac{\sum_{|a-b|=i} D(a,b)}{n_i}$$

where n_i is the number of site pairs having true geographic distance $i = 1, \ldots, n-1$. So $f(i)$ represents the average distance observed between all site pairs seperated by true geographic distance i.

Figure 4.3 shows that Hellinger's distance is clearly monotone. Another type of analysis considers how much of the calculated Hellinger distance can be explained by

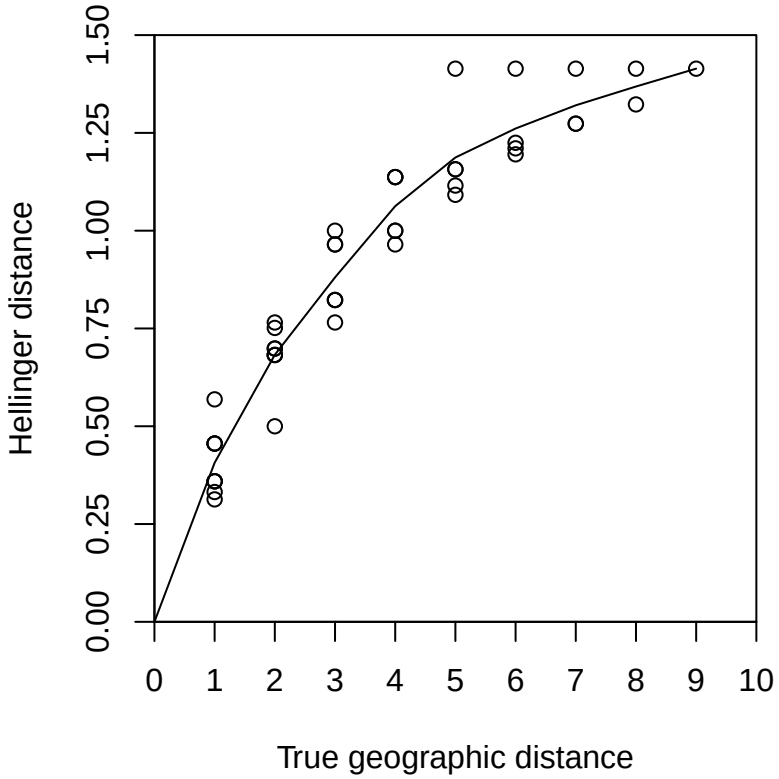

Figure 4.3: Hellinger Distance for Artificial Gradient Data

the true geographic distance. [Legendre and Gallagher, 2001] refers to this measure as R^2 since it has computation similar to the coefficient of determination used in linear regression. Let $\bar{d}$ reprent the overall mean distance between all (a,b) pairs. Note that out of n sites, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ pairings between sites. So the overall average distance can be defined as

$$\bar{d} = \frac{\sum\limits_{(a,b)} D(a,b)}{\binom{n}{2}}$$

For a specific ecological distance D, this value is given by

$$R^2 = \frac{\sum\limits_{(a,b)} \big(f(i) - \bar{d}\big)^2}{\sum\limits_{(a,b)} \big(D(a,b) - \bar{d}\big)^2}$$

In the case of Hellinger's distance, we obtain $R^2 = 0.9418$. Let's compare these results with other popular ecological distances by looking at Figure 4.4.

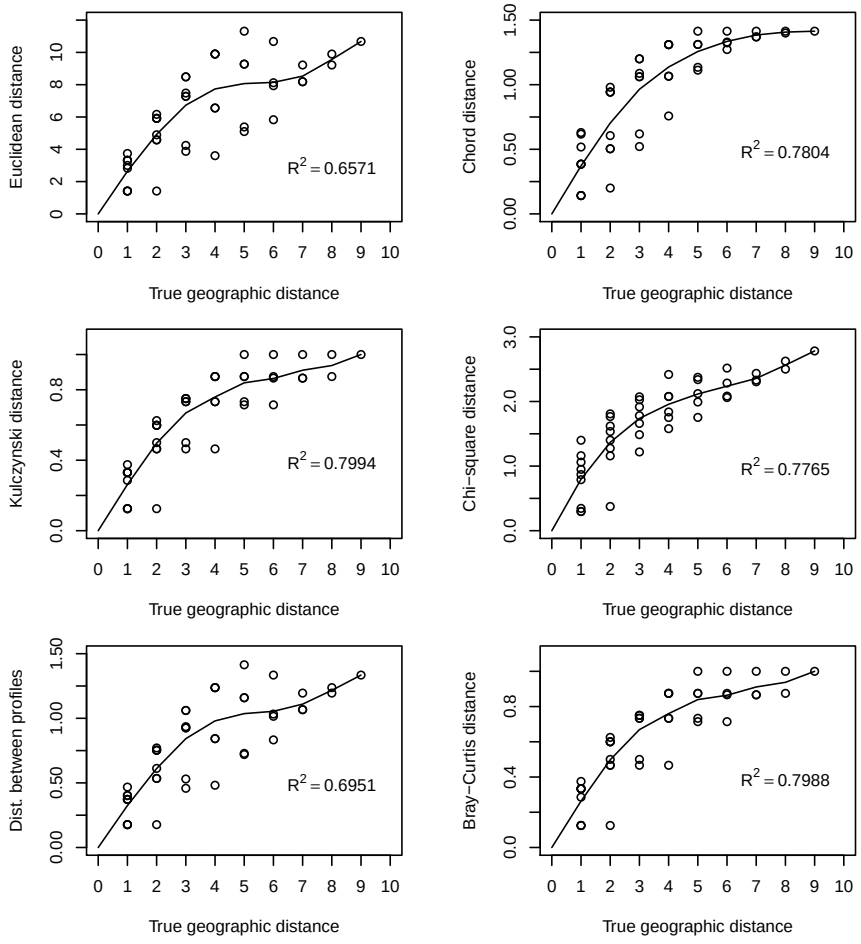

Figure 4.4: Popular Ecological Distances for Artificial Gradient Data

It is clear from Figure 4.4 that the Euclidean and Species Profile distances

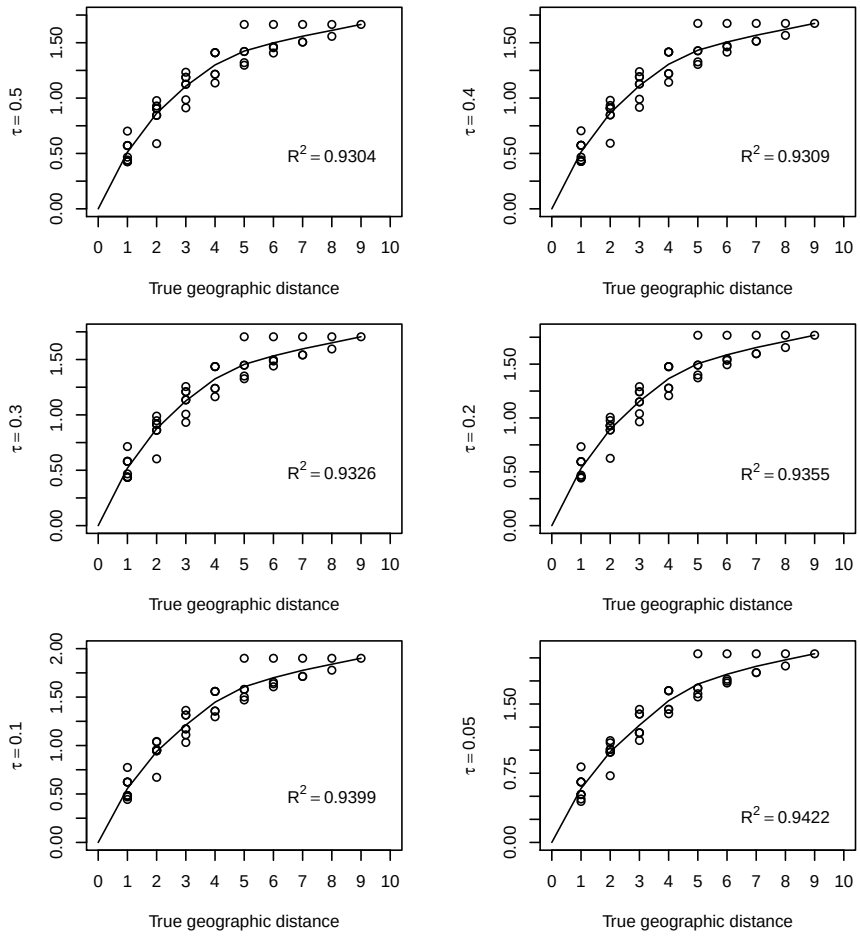

Figure 4.5: Symmetric GKL Distances for Artificial Gradient Data

would be poor choices for this data set. The Kulczynski distance outperforms the remaining distances, but is inferior to Hellinger's distance. In Figure 4.5, we consider an ecological distance based on our symmetric generalized Kullback-Leibler distance for different values of the tuning parameter τ. It appears that the R^2 values are increasing for this distance as τ gets closer to zero. At $\tau = 0.05$, an $R^2 = 0.9422$ is achieved which is better than that of Hellinger ($R^2 = 0.9418$) so it appears the symmetric generalized Kullback-Leibler distance is a good choice.

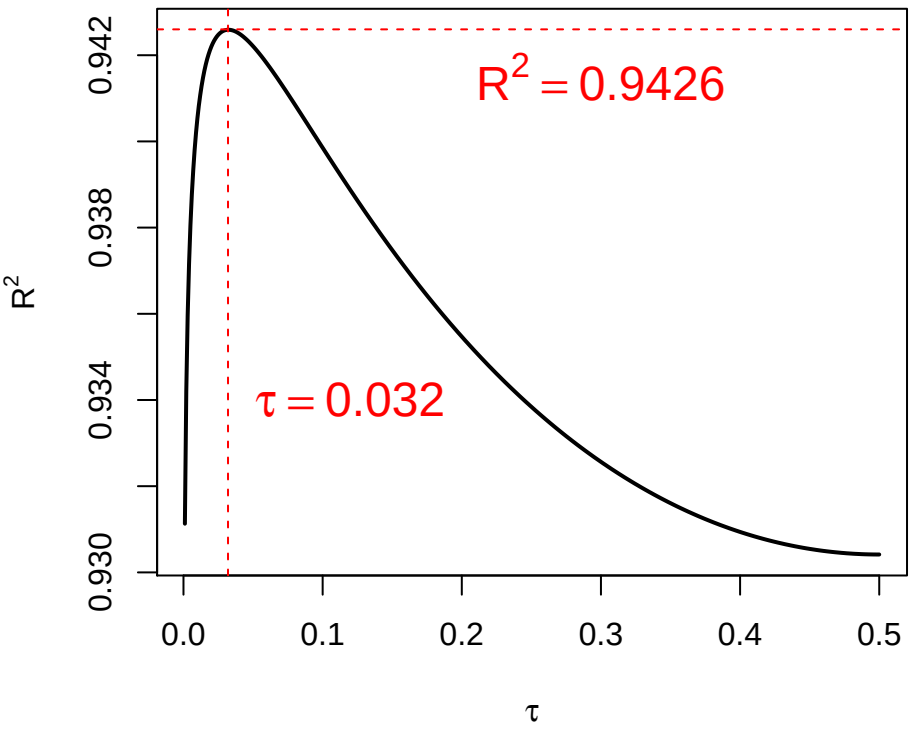

Figure 4.6: Symmetric GKL R^2 values for Artificial Gradient Data

Furthermore, in Figure 4.6 a plot is shown for the R^2 values for $\tau \in [0, 0.5]$. A maximum value of $R^2 = 0.9426$ is acheived when $\tau = 0.032$.

Similarly, the ecological distances can be computed using either the symmetric blended weight Hellinger distance (SBWHD) or symmetric blended weight chi-square distance (SBWCS). Figures 4.7 and 4.8 show the R^2 values given by using each distance.

Overall it appears that R^2 values for SBWHD are better for larger values of τ, and it achieves a maximum of 0.9424 when $\tau = 0.402$. Also the R^2 values for SBWCS are

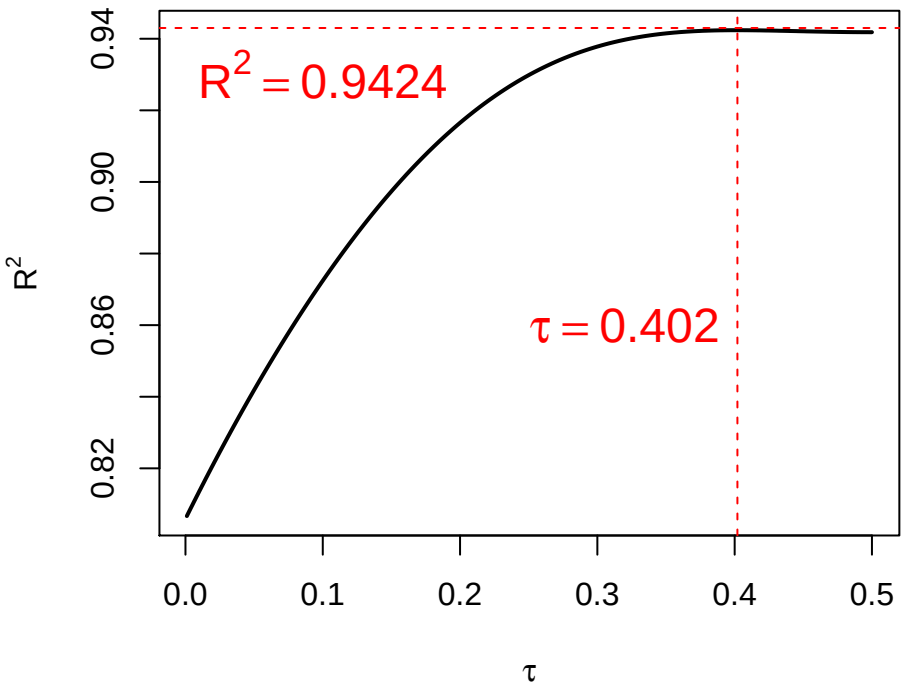

Figure 4.7: Symmetric BWHD R^2 values for Artificial Gradient Data

better for lower values of τ, and it achieves a maximum of 0.9051 when $\tau = 0.064$.

Now consider the a second artificial gradient data set from [Legendre and Gallagher, 2001] shown in Table 4.4:

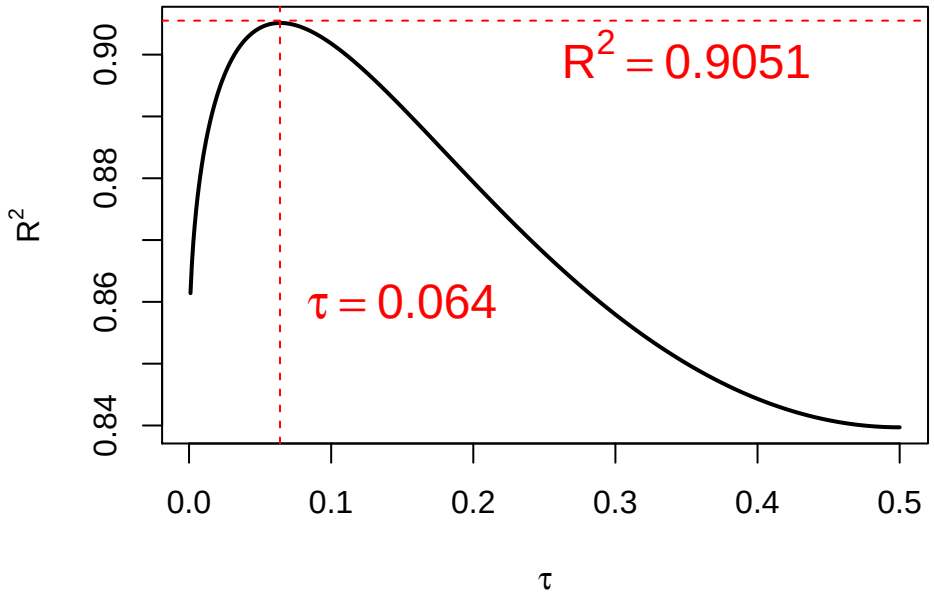

Figure 4.8: Symmetric BWCS R^2 values for Artificial Gradient Data

Figure 4.9 shows the R^2 values for SGKL, SBWHD, and SBWCS for values of $\tau \in [0, 0.5]$. The graphs are similar to those for the first data sets. Both SGKL and SBWCS appear to do better for τ close to 0, while SBWHD appears to do better when τ is closer to 0.5. The R^2 value for each ecological distance was calculated and recorded in Table 4.6. This table includes the values of τ for which SGKL, SBWHD, and SBWCS achieve a maximum R^2 value.

Compared to other popular distances that are being used in ecology, it seems like all three of the distances proposed in the previous chapter perform well for species composition data. Specifically, for both data sets, it has been shown that two of the three distances appear to out-perform Hellinger's distance.

Table 4.5: Artificial dataset with the abundance of 9 species for 19 sites

	Species								
Sites	1	2	3	4	5	6	7	8	9
1	7	1	0	0	0	0	0	0	0
2	4	2	0	0	0	1	0	0	0
3	2	4	0	0	0	1	0	0	0
4	1	7	0	0	0	0	0	0	0
5	0	8	0	0	0	0	0	0	0
6	0	7	1	0	0	0	0	0	0
7	0	4	2	0	0	0	2	0	0
8	0	2	4	0	0	0	1	0	0
9	0	1	7	0	0	0	0	0	0
10	0	0	8	0	0	0	0	0	0
11	0	0	7	1	0	0	0	0	0
12	0	0	4	2	0	0	0	3	0
13	0	0	2	4	0	0	0	1	0
14	0	0	1	7	0	0	0	0	0
15	0	0	0	8	0	0	0	0	0
16	0	0	0	7	1	0	0	0	0
17	0	0	0	4	2	0	0	0	4
18	0	0	0	2	4	0	0	0	1
19	0	0	0	1	7	0	0	0	0

Table 4.6: Results for Second Data Set

Distance	R^2
Chord	0.82709
Euclidean	0.63736
Chi-Square Distance	0.65633
Species Profiles	0.66882
Hellinger	0.95433
Bray Curtis	0.85537
Kulczynski	0.85623
SGKL with $\tau = 0.015$	0.95623
SBWHD with $\tau = 0.348$	0.95618
SBWCS with $\tau = 0.043$	0.91608

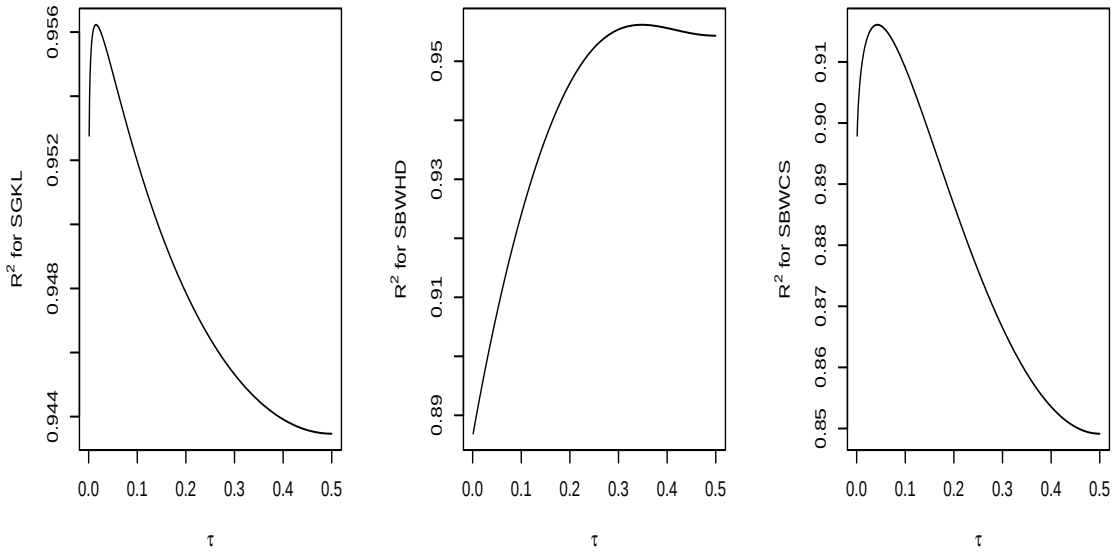

Figure 4.9: Symmetric GKL, BWHD, and BWCS R^2 values for Artificial Gradient Data

Chapter 5

Future Work

5.1 Improving Results

In future work, we plan to further investigate the properties of the proposed class of symmetric distances. It is our belief that all can be approximated by a chi-square statistic. Once these results have been established, the goal will be to develop a new criteria that will help in choosing distances in goodness-of-fit testing when symmetric distances are needed. These distances should also perform well in the context of parameter estimation.

The next goal will be to investigate the properties of the new weight function introduced in Chapter 3 for symmetric distances and applications. In parameter estimation, it made sense to compare distances with the likelihood disparity since it is the best under normal conditions. In goodness of fit testing, we suggested that symmetric chi-square distance, GKL with $\tau = 0.5$, or Hellinger's distance may be good choices for weights since we are no longer focused on parameter estimation. It is our intuition that the weight function will be able to give us more information about how symmetric distances handle large discrepancies. It may also be necessary to consider

a new residual adjustment function and develop new methods for determining the robustness and efficiency of symmetric disparity measures.

Lastly, one of the biggest goals in the future will be to develop a general approach to choosing distances when analyzing species composition data. According to [Legendre and Gallagher, 2001], "theoretical criteria are not known at the moment that would allow one to select the best distance function for any specific situation." However, the rare species problem in ecology is very similar to the empty cell problem in parameter estimation (i.e. extreme outliers). There is also theory we believe can be used from the empty cell problem to improve distance calculations when a species is absent from both sites (i.e. double zeros). Ultimately, it seems like there are plenty of applications where symmetry is needed that could also benefit from a statistical distance approach. The next section includes an application in image segmentation. In future work, similar applications will be explored in greater detail.

5.2 Image Segmentation Problem

The process of segmenting or partitioning an image into disjoint regions is very important in many applications [Sandhu et al., 2008]. The goal of this type of segmentation is typically to distinguish between objects in the picture. In certain applications, such as medical image analysis, this task can become difficult due to distortions of the image. The picture on the left in Figure 5.1 clearly shows a digital image consisting of four distinct regions. Segmentation aims to properly detect the "borders" that separate each region. Note that the when images are distorted, it can become very difficult to distinguish between regions. It is very important to have segmentation procedures that will perform well even when the image is not clear. Although there are many current methods of segmentation, this section will focus on

histogram-based segmentation.

In computer vision, a digital image is made up of tiny elements of the original image called pixels. The intesity of the pixel varies throughout the image. This section will consider the grayscale intensity system, which assigns each pixel a value between 0 and 1. In this system, 0 represents the color black and 1 represents the color white. Also this means that a pixel with intensity close to 0 will be represented by a dark-gray image and a pixel with intensity close to 1 will be represented by a light-gray image. In the analysis at the end of this section, we use the EBI package in statistical software R [Pau et al.,] to convert images to intensity values and vice versa.

Figure 5.1: Original image (left) and Image Boundary (right)

In histogram-based image segmentation, collections of pixels are compared to each other using a distance measure [Arifin and Asano, 2006]. Consider a specific pixel p in a digital image. A specified number of pixels are gathered from the left and right of p. A histogram is created for each collection of pixels and the relative frequencies are calculated. The distance measure is then used to quantify the discrancy between the two collections. Similarly, we can consider gathering a collection of pixels above

p and below p, obtain two more histograms, and then calculate the distance between them. Figure 5.2 shows an example of how pixels are gathered from all four directions to create the four histograms. The final distance we use for that pixel will be the average of the two distances. It should be noted that our final distance should be between 0 and 1, so that our grayscale intensity value can be interpreted correctly.

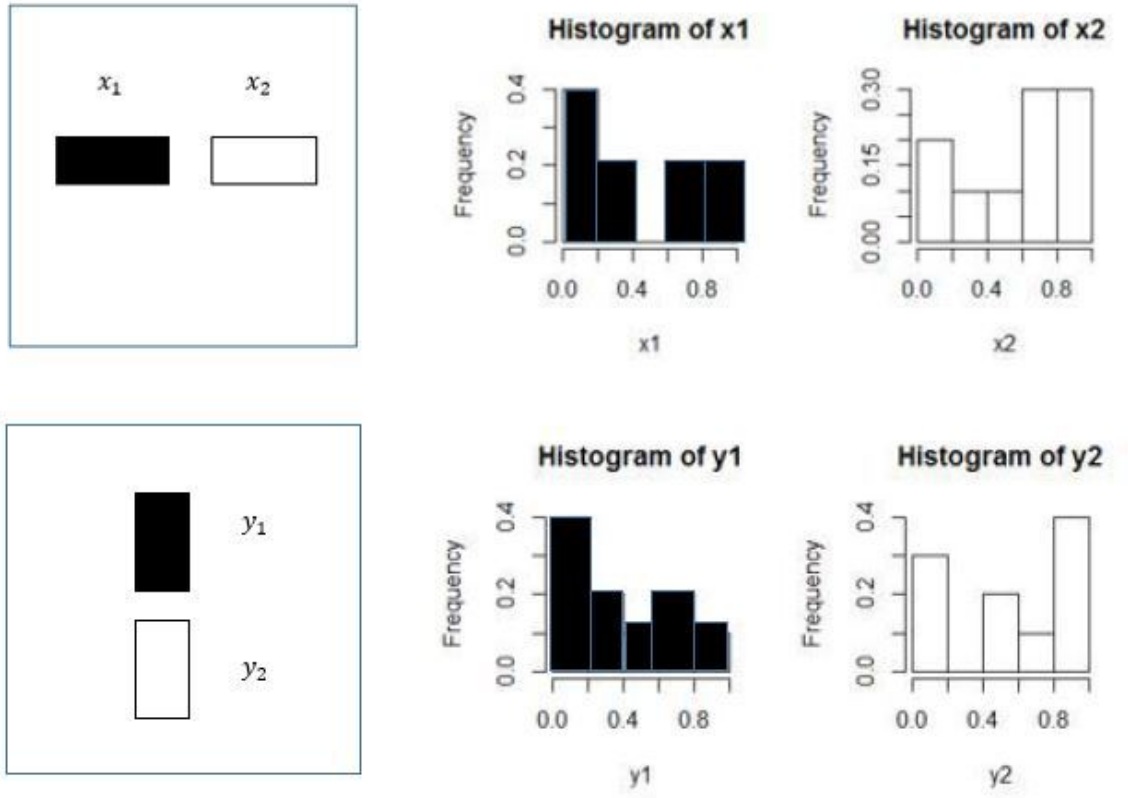

Figure 5.2: Example of Histogram-based image segmentation

Several of the distances used in this paper have been used in the past. The Kullback-Leibler distance is used in [Schroff et al., 2006] and the Hellinger distance in [V. Gonzalez-Castro and Alegre, 2013]. It is mentioned in [Schroff et al., 2006] that the Kullback-Leibler distance is not ideal since it is not symmetric. Several generalized distances are even proposed in [He et al., 2003], [Sandhu et al., 2008],

and [Basseville, 2010]. However, the distances introduced in this paper have not been used yet in the field of image segmentation. Before using our new symmetric distances, we must first address the fact that our distances are not between 0 and 1.

It is clear that SGKL_τ, SBWHD_τ, and SBWCS_τ have the same range as GKL_τ, BWHD_τ, and BWCS_τ, respectively. It is proven in [Park and Basu, 2003] that GKL_τ is bounded above for $0 < \tau < 1$ and

$$\mathrm{GKL}_\tau(d, f_\theta) \leq \frac{1}{\tau} \log\left(\frac{1}{1-\tau}\right) + \frac{1}{1-\tau} \log\left(\frac{1}{\tau}\right)$$

The following theorems also give us upper bounds for the BWHD and BWCS:

Theorem 5.2.1. *The blended weight Hellinger distance is bounded above for* $0 < \tau < 1$. *Specifically, for all values of* y

$$\mathrm{BWHD}_\tau(d, f_\theta) \leq \frac{1}{2}\Big[\frac{1}{(1-\tau)^2} + \frac{1}{\tau^2}\Big]$$

Proof. See Appendix B □

Theorem 5.2.2. *The blended weight chi-square distance is bounded above for* $0 < \tau < 1$. *Specifically, for all values of* y

$$\mathrm{BWCS}_\tau(d, f_\theta) \leq \frac{1}{2}\Big[\frac{1}{1-\tau} + \frac{1}{\tau}\Big] = \frac{1}{2\tau(1-\tau)}$$

Proof. See Appendix B □

Using these upper bounds we can obtain new versions of our distances that will be appropriate for this application:

$$\mathrm{SGKL}_\tau = \frac{\mathrm{GKL}_\tau(d, f_\theta) + \mathrm{GKL}_\tau(f_\theta, d)}{2\left[\frac{1}{\tau}\log\left(\frac{1}{1-\tau}\right) + \frac{1}{1-\tau}\log\left(\frac{1}{\tau}\right)\right]}$$

$$\mathrm{SBWHD}_\tau = \frac{\mathrm{BWHD}_\tau(d, f_\theta) + \mathrm{BWHD}_\tau(f_\theta, d)}{\left[\frac{1}{(1-\tau)^2} + \frac{1}{\tau^2}\right]}$$

$$\mathrm{SBWCS}_\tau = \frac{\mathrm{BWCS}_\tau(d, f_\theta) + \mathrm{BWCS}_\tau(f_\theta, d)}{\tau(1-\tau)}$$

The image we will use for analysis is one of the most famous pictures in image processing. The first picture in Figure 5.3 was taken in 1972 of the model Lena Soderber [Gonzalez and Woods, 2008]. The second picture represents the image after using histogram-based segmentation. If our distances are appropriate, we would like for our results to at least by comparable with that of Hellinger's distance.

Figures 5.4, 5.5, and 5.6 show the results of histogram-based image segmentation using different values of $0 < \tau < 1$ for SGKL_τ, SBWHD_τ, and SBWCS_τ, respectively. The SBWHD_τ and SBWCS_τ appear to perform very well for most values of τ. The SGKL_τ distance did not perform as well, but we believe the results can be improved in the future.

Figure 5.3: Grayscale Lena (left) and Lena after Image Segmentation based on Hellinger Distance (right)

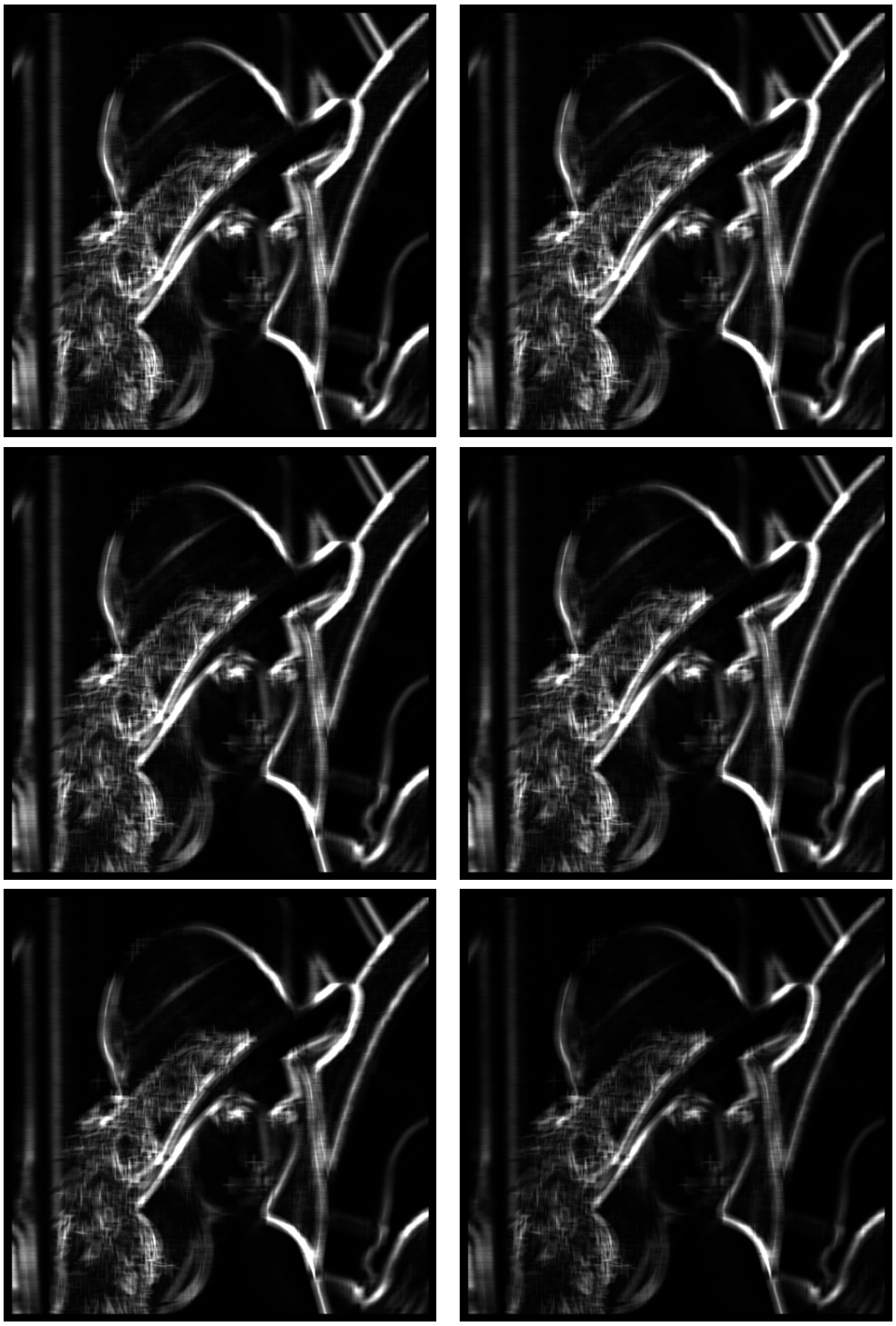

Figure 5.4: SGKL (From top-left to bottom-right): $\tau = 0.5$, $\tau = 0.4$, $\tau = 0.3$, $\tau = 0.2$, $\tau = 0.1$, and $\tau = 0.01$

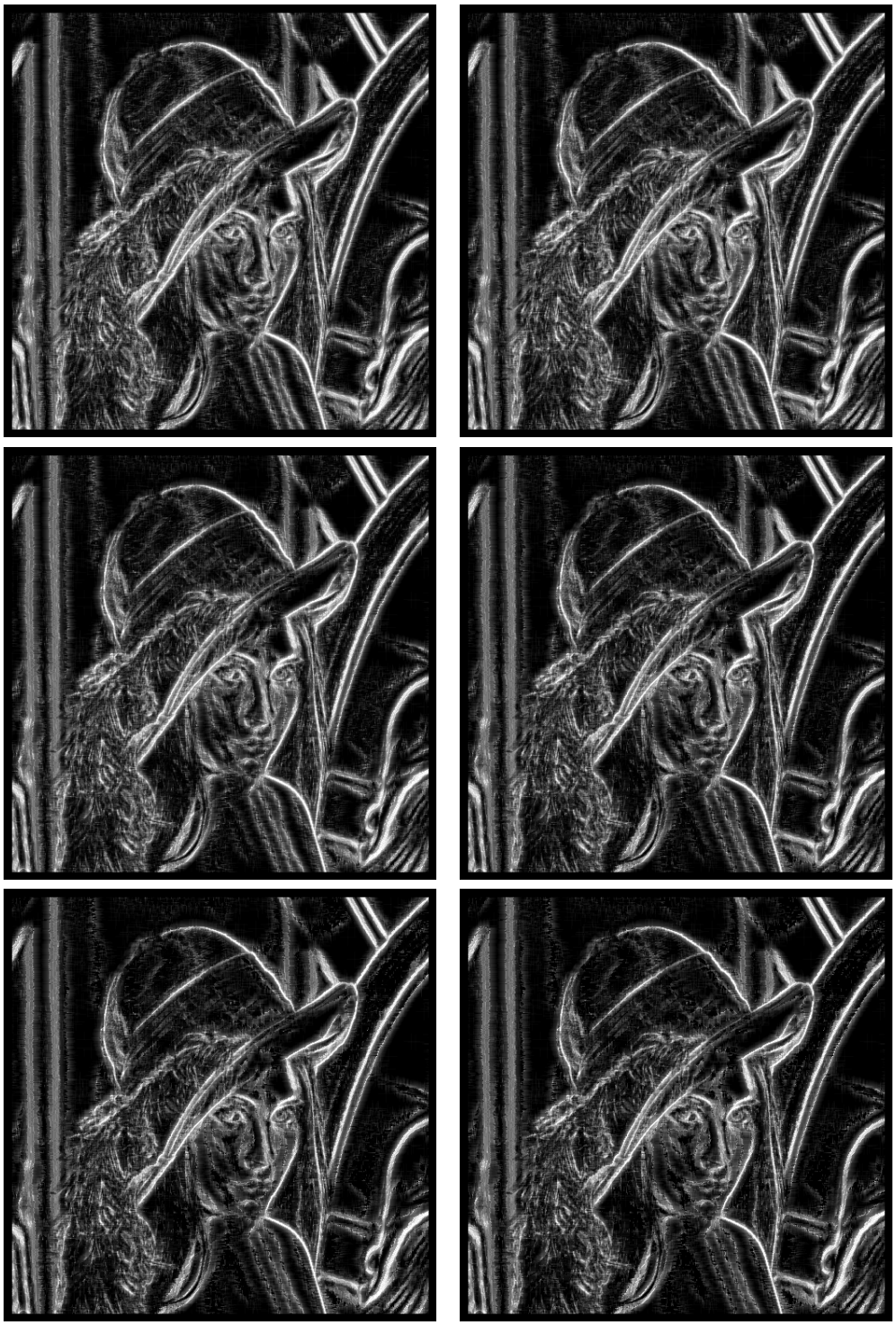

Figure 5.5: SBWHD (From top-left to bottom-right): $\tau = 0.5$, $\tau = 0.4$, $\tau = 0.3$, $\tau = 0.2$, $\tau = 0.1$, and $\tau = 0.01$

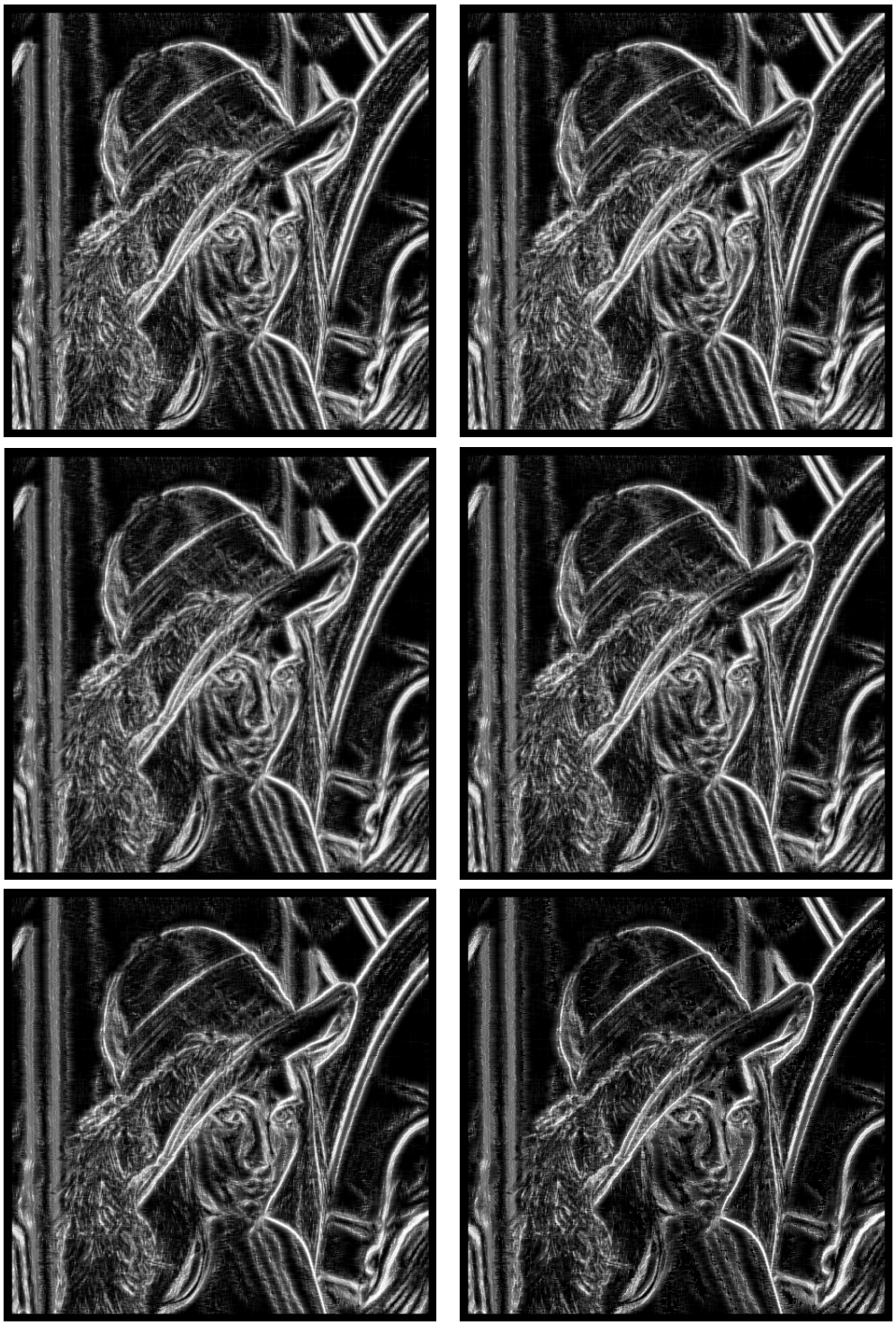

Figure 5.6: SBWCS (From top-left to bottom-right): $\tau = 0.5$, $\tau = 0.4$, $\tau = 0.3$, $\tau = 0.2$, $\tau = 0.1$, and $\tau = 0.01$

Appendices

Appendix A

Summarizing Distances

As a reference we summarize some of the common disparities by specifying the G, RAF, and weight functions. We also mention some key relationships and properties for each of the following:

1. LD	2. KL	3. HD
4. PCS	5. NCS	6. SCS
7. GKL_τ	8. $BWHD_\tau$	9. $BWCS_\tau$
10. $SGKL_\tau$	11. $SBWHD_\tau$	12. $SBWCS_\tau$

Before we summarize the properties of popular distances, recall the following concepts from Chapter 2:

Residual Functions

Let $f_\theta(x)$ represent the model density evaluated at x and let $d(x)$ represent the empirical density evaluated at x.

Pearson's Residual: $\delta_P(x) = \frac{d(x)-f_\theta(x)}{f_\theta(x)}$.

Neyman's Residual: $\delta_N(x) = \frac{d(x)-f_\theta(x)}{d(x)}$.

Combined Residual:

$$\delta_C(x) = \begin{cases} \delta_P(x) & : \text{when } d(x) \leq f_\theta(x) \\ \delta_N(x) & : \text{when } d(x) > f_\theta(x). \end{cases}$$

Symmetric Residual: $\delta_S(x) = \frac{d(x)-f_\theta(x)}{d(x)+f_\theta(x)} = \frac{2\delta_P(x)}{1-\delta_P(x)}$.

Desirable Properties

Robust estimators: If $\rho_G(\cdot)$ is bounded and $G'(\infty)$ is finite, then the asymptotic breakdown point of the minimum disparity estimators is at least $\frac{1}{2}$. Recall the breakdown point of an estimator is the proportion of incorrect observations an estimator can handle before giving an incorrect result. By definition the maximum breakdown point is 0.5 and estimators that achieve such a breakdown point are called **robust** or **resistant**. The property: $\lim_{\delta_P \to +\infty} w_c(\delta_P) = \lim_{\delta_P \to +\infty} \frac{A(\delta_P)}{\delta_P} = 0$ is a necessary condition for robustness.

Inlier robust estimators: $\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P} = 0$ is a necessary condition for inlier robustness.

Second Order Efficiency: $A''(0) = 0$ is a necessary condition for second order efficiency.

Check order of $G(\delta_P)$: Implode $< O(\delta_P) \leq$ Reasonable $\leq O(\delta_P \log(\delta_P)) <$ Explode .

If $G(-1)$ and $G'(\infty)$ are finite and $A(\delta_P) \geq 0$ for $\delta_P > 0$, the disparity $\rho_G(\cdot)$ is bounded above by $G(-1) + G'(\infty)$.

1. Likelihood Disparity (LD)

$$\mathrm{LD}(d(y), f_\theta(y)) = \sum_{\forall y} \left[d(y) \log \frac{d(y)}{f_\theta(y)} - d(y) + f_\theta(y) \right]$$

$$G_{\mathrm{LD}}(\delta_P) = (\delta_P + 1) \log \Big(\delta_P + 1 \Big) - \delta_P$$

$$G_{\mathrm{LD}}(-1) = 1$$

$$G_{\mathrm{LD}}(\infty) = \infty$$

$$A_{\mathrm{LD}}(\delta_P) = \delta_P$$

Robust? $\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	Inlier Robust? $\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	Efficiency? $A''(0)$
1	1	0
No	No	Second Order

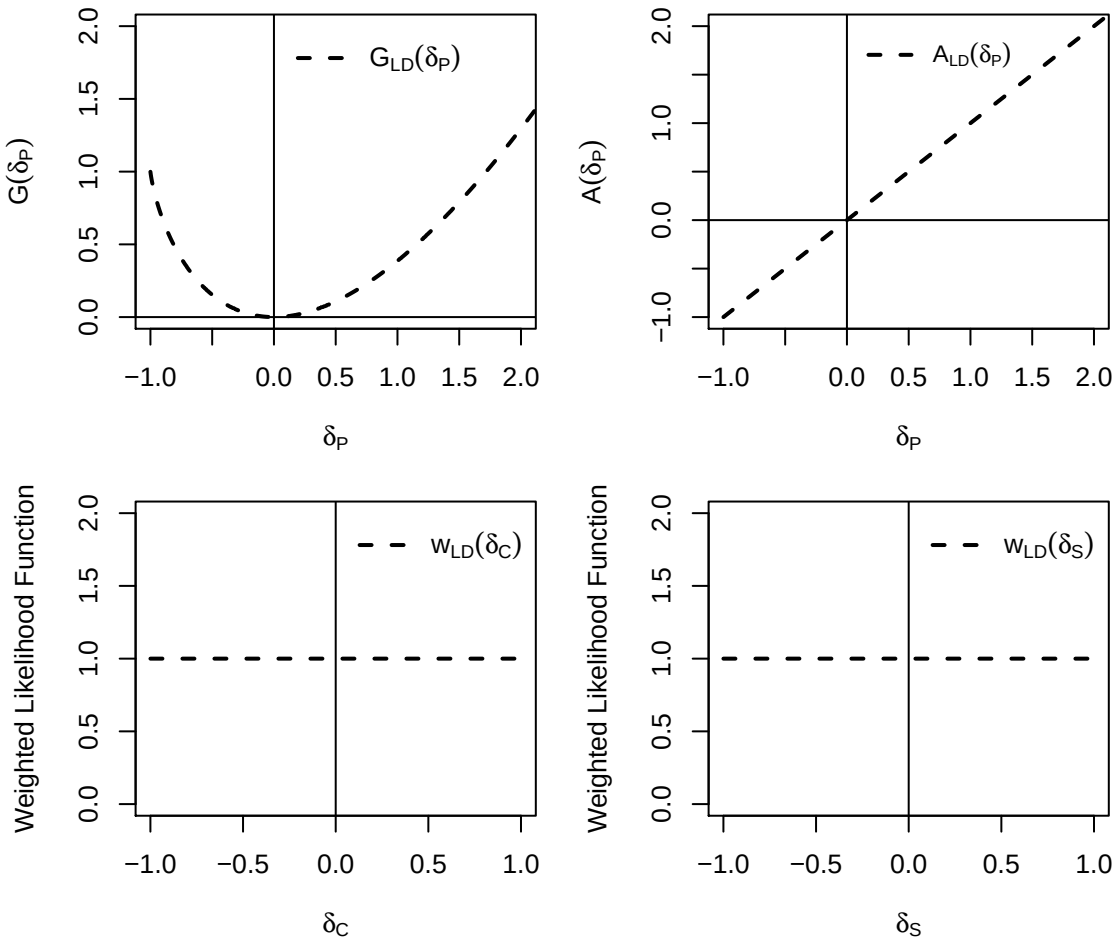

Figure A.1: Likelihood

2. Kullback-Leibler (KL)

$$\mathrm{KL}(d(y), f_\theta(y)) = \sum_{\forall y} \left[f_\theta(y) \log \frac{f_\theta(y)}{d(y)} + d(y) - f_\theta(y) \right]$$

$$G_{\mathrm{KL}}(\delta_P) = -\log\Big(\delta_P + 1\Big) + \delta_P$$

$$G_{\mathrm{KL}}(-1) = \infty$$

$$G_{\mathrm{KL}}(\infty) = \infty$$

$$A_{\mathrm{KL}}(\delta_P) = \log\Big(\delta_P + 1\Big)$$

Robust? $\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	Inlier Robust? $\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	Efficiency? $A''(0)$
0	∞	-1
Yes	No	First Order

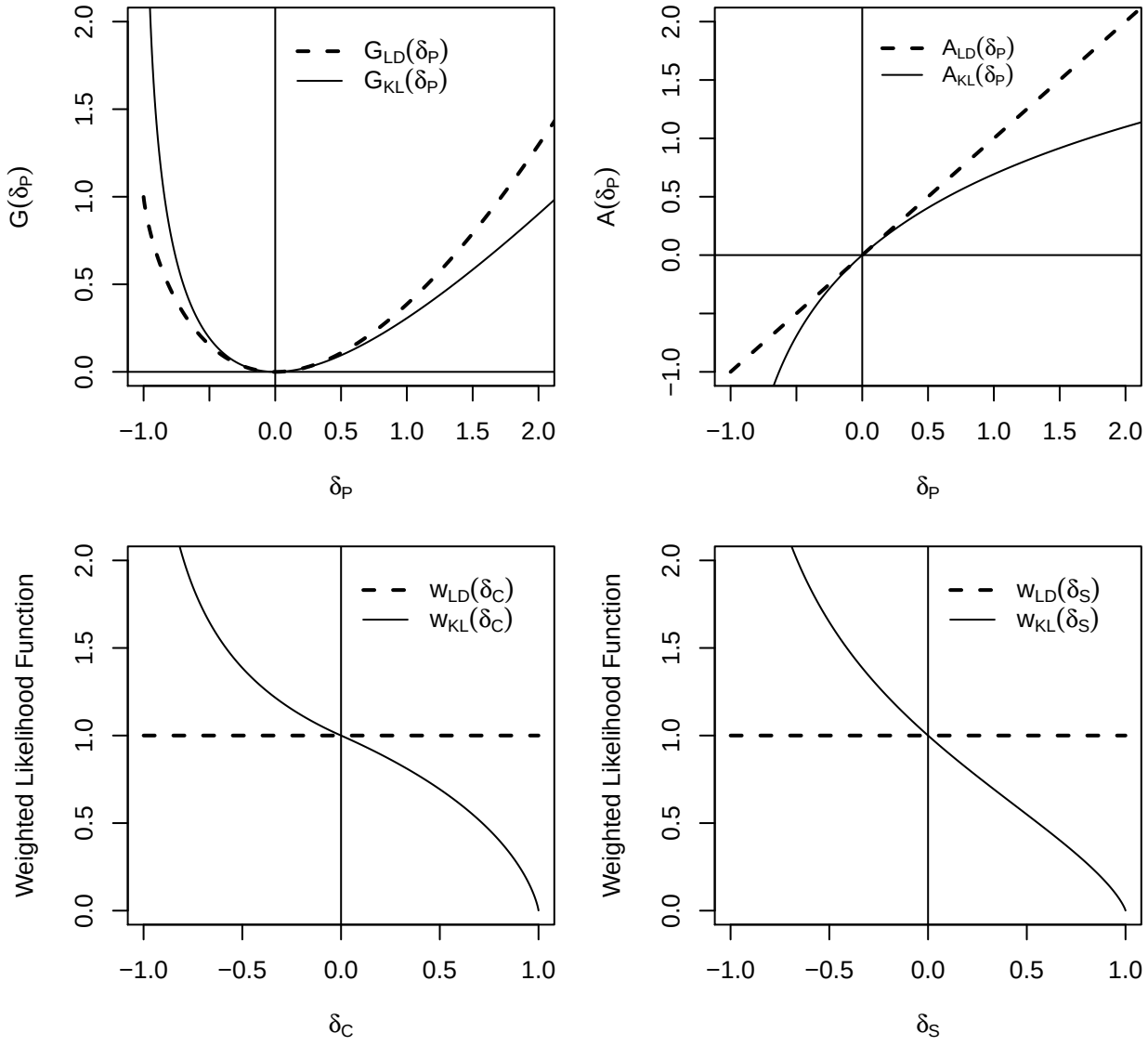

Figure A.2: Kullback-Leibler

3. Hellinger Distance (HD)

$$\mathrm{HD}(d(y), f_\theta(y)) = 2\sum_{\forall y}\left[\sqrt{d(y)} - \sqrt{f_\theta(y)}\right]^2$$

$$G_{\mathrm{HD}}(\delta_P) = 2\Big(\sqrt{\delta_P + 1} - 1\Big)^2$$

$$G_{\mathrm{HD}}(-1) = 2$$

$$G_{\mathrm{HD}}(\infty) = \infty$$

$$A_{\mathrm{HD}}(\delta_P) = 2\Big(\sqrt{\delta_P + 1} - 1\Big)$$

Robust?	Inlier Robust?	Efficiency?
$\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	$\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	$A''(0)$
0	2	$-\frac{1}{2}$
Yes	No	First Order

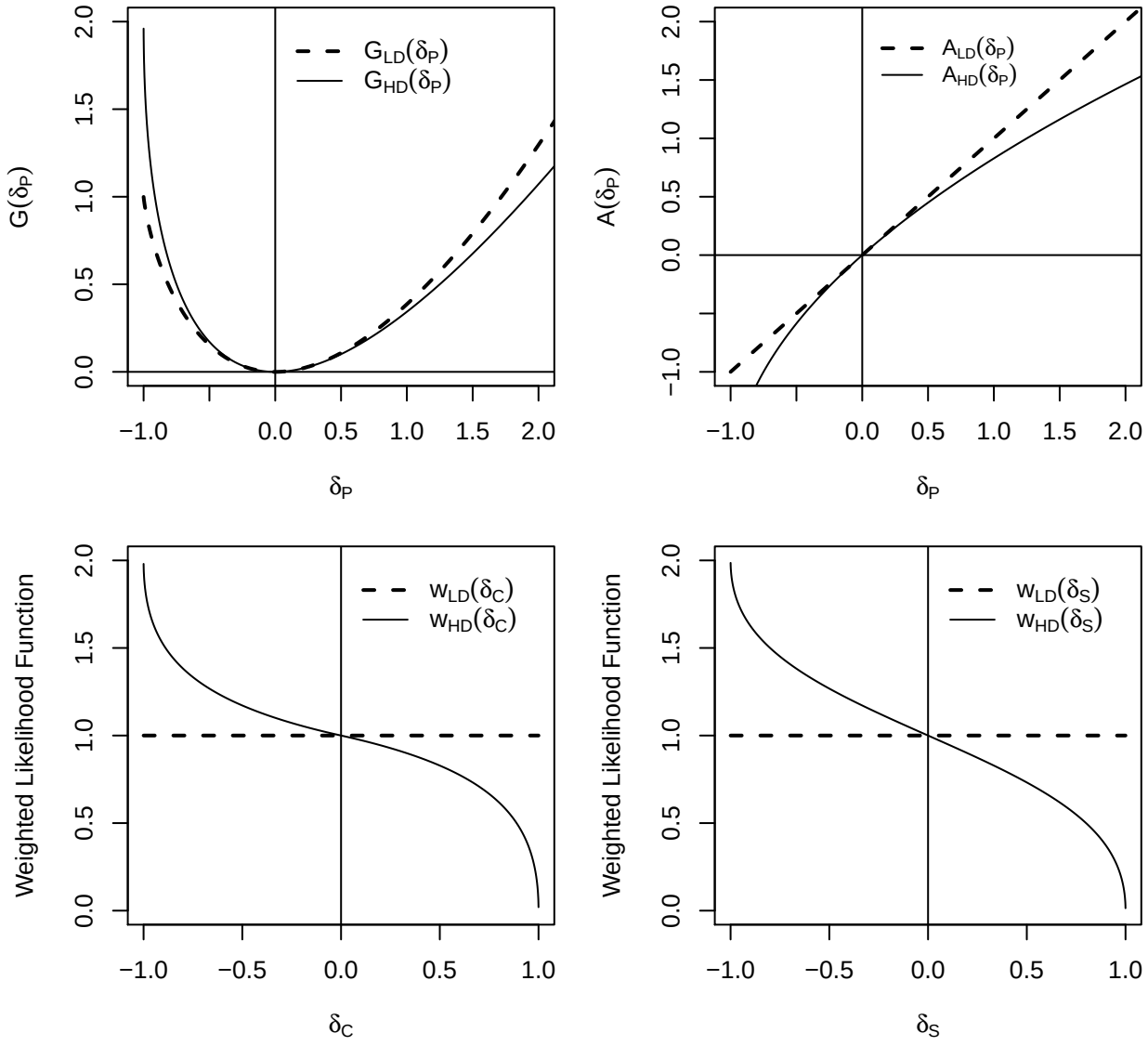

Figure A.3: Hellinger

4. Pearson chi-square (PCS)

$$\mathrm{PCS}(d(y), f_\theta(y)) = \frac{1}{2}\sum_{\forall y}\left[\frac{\Big(d(y) - f_\theta(y)\Big)^2}{f_\theta(y)}\right]$$

$$G_{\mathrm{PCS}}(\delta_P) = \frac{1}{2}\delta_P^2$$

$$G_{\mathrm{PCS}}(-1) = \frac{1}{2}$$

$$G_{\mathrm{PCS}}(\infty) = \infty$$

$$A_{\mathrm{PCS}}(\delta_P) = \delta_P + \frac{1}{2}\delta_P^2$$

Robust?	Inlier Robust?	Efficiency?
$\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	$\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	$A''(0)$
∞	$-\frac{1}{2}$	0
No	No	Second Order

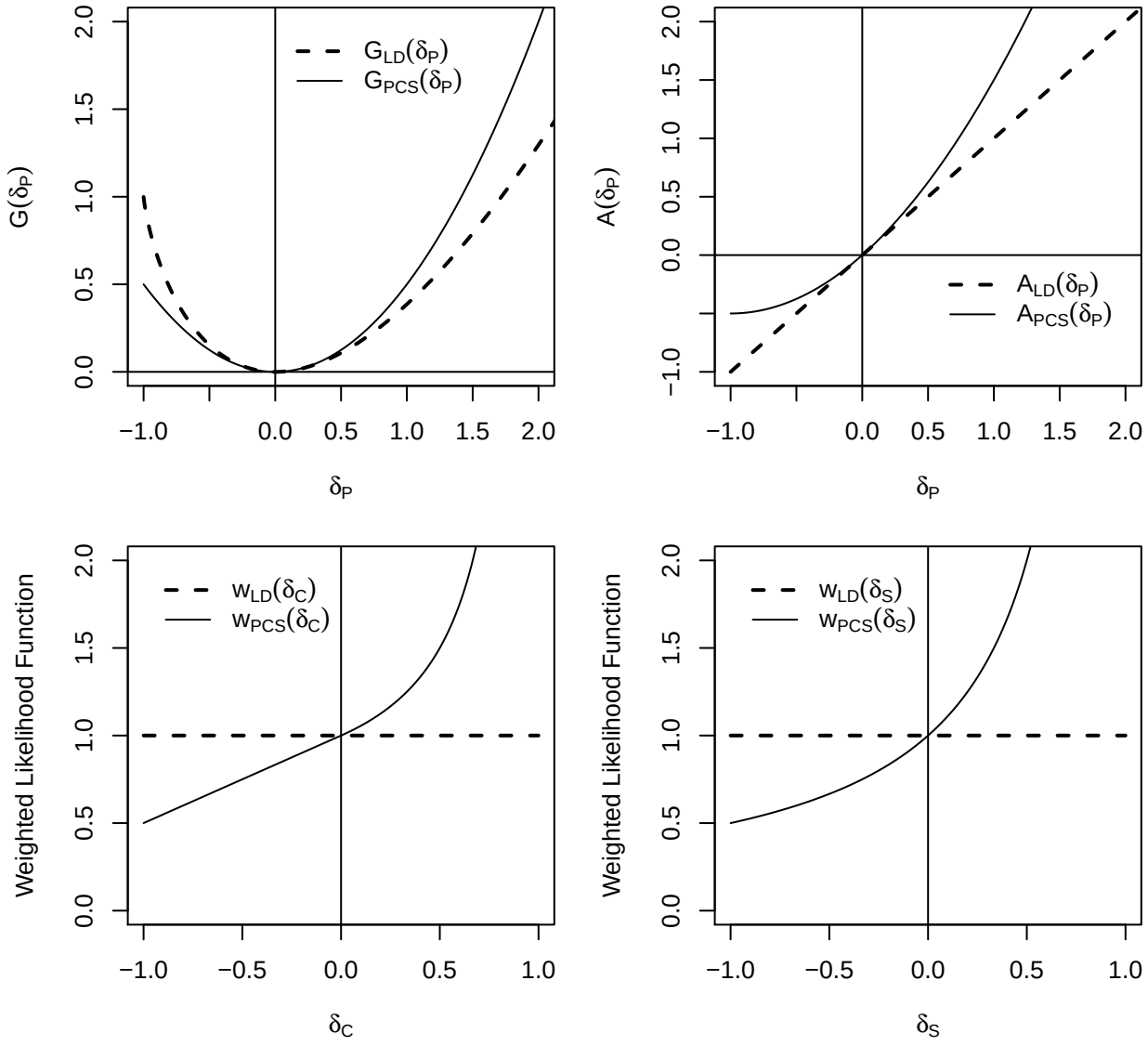

Figure A.4: Pearson Chi-Square

5. Neyman Chi-Square (NCS)

$$\mathrm{NCS}(d(y), f_\theta(y)) = \frac{1}{2}\sum_{\forall y}\left[\frac{\Big(d(y) - f_\theta(y)\Big)^2}{d_\theta(y)}\right]$$

$$G_{\mathrm{NCS}}(\delta_P) = \frac{\delta_P^2}{2\Big(\delta_P + 1\Big)}$$

$$G_{\mathrm{NCS}}(-1) = \infty$$

$$G_{\mathrm{NCS}}(\infty) = \infty$$

$$A_{\mathrm{NCS}}(\delta_P) = \frac{\delta_P}{\delta_P + 1}$$

Robust?	Inlier Robust?	Efficiency?
$\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	$\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	$A''(0)$
1	$-\infty$	-2
No	No	First Order

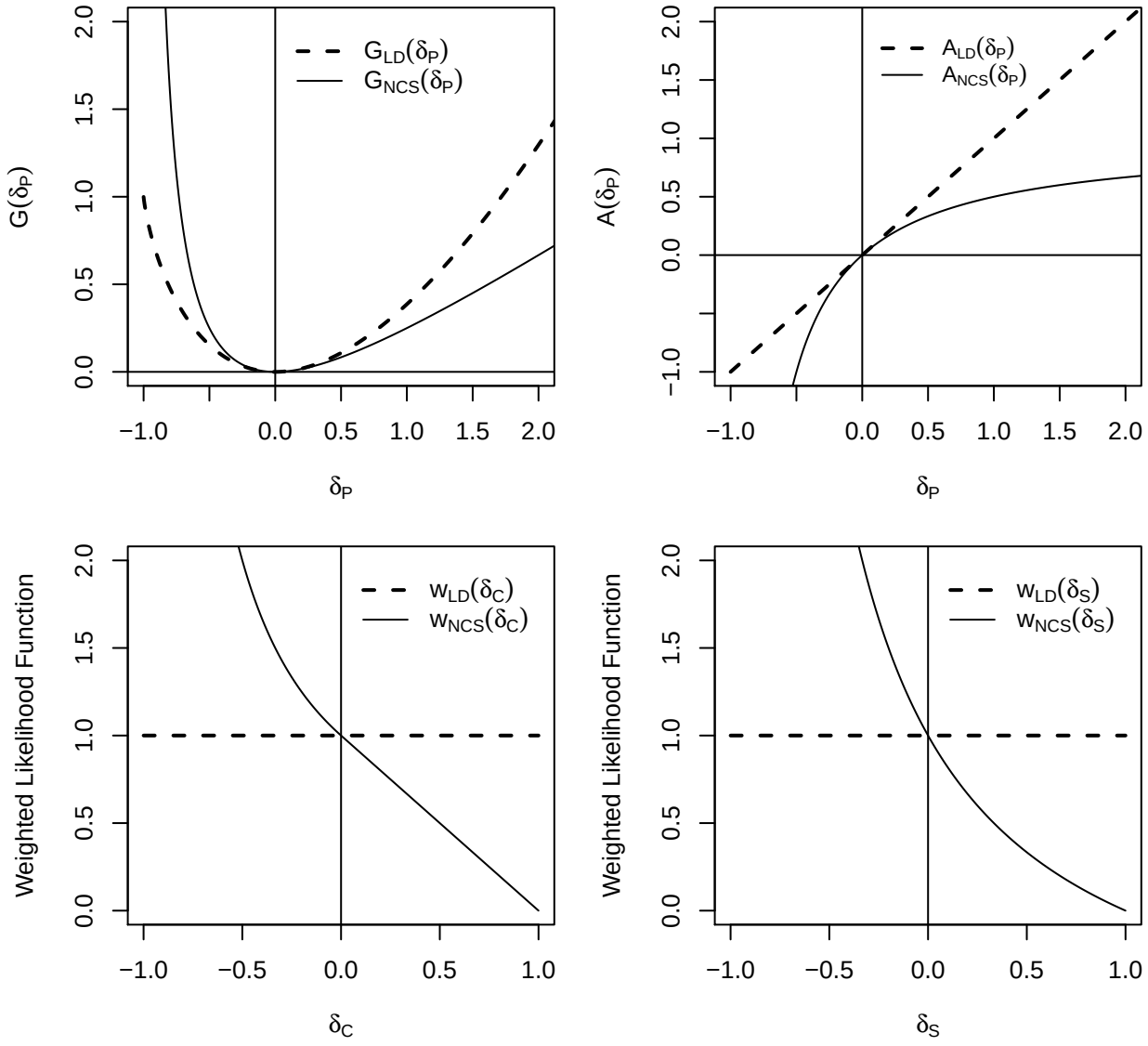

Figure A.5: Neyman Chi Square

6. Symmetric Chi-Square (SCS)

$$\mathrm{SCS}(d(y), f_\theta(y)) = \frac{1}{2} \sum_{\forall y} \left[\frac{\Big(d(y) - f_\theta(y)\Big)^2}{d(y) + f_\theta(y)} \right]$$

$$G_{\mathrm{SCS}}(\delta_P) = \frac{\delta_P^2}{\delta_P + 2}$$

$$G_{\mathrm{SCS}}(-1) = 1$$

$$G_{\mathrm{SCS}}(\infty) = \infty$$

$$A_{\mathrm{SCS}}(\delta_P) = \frac{3\delta_P + 4}{\Big(\delta_P + 2\Big)^2}$$

Robust? $\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	Inlier Robust? $\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	Efficiency? $A''(0)$
3	-1	$-\frac{1}{2}$
No	No	First Order

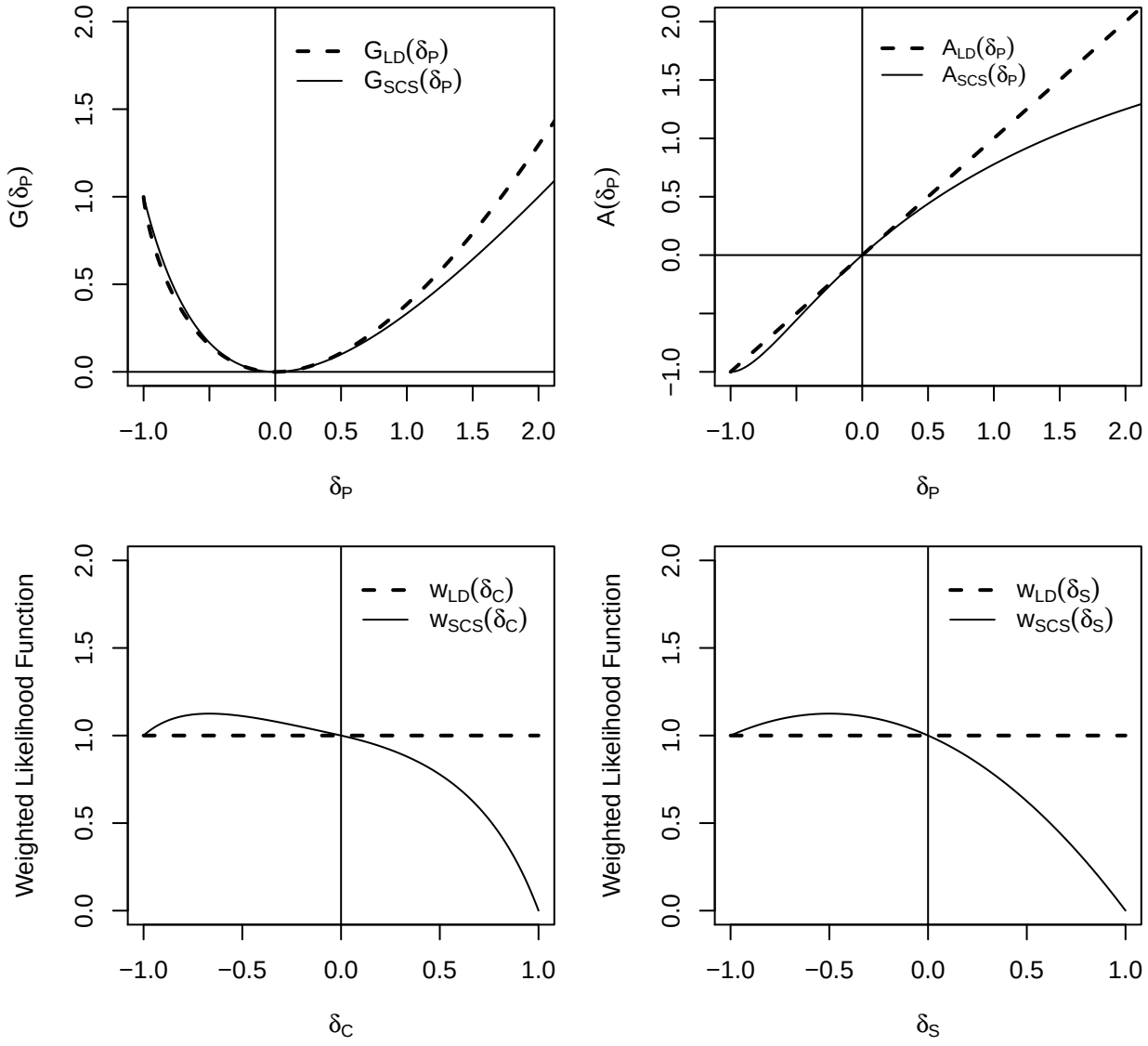

Figure A.6: Symmetric Chi Square

7. Generalized Kullback-Leibler (GKL)

For $\tau \in [0,1]$ we have the following

$$\mathrm{GKL}(d(y), f_\theta(y)) = \sum_{\forall y} \left[\frac{d(y)}{1-\tau} \log\Big(\frac{d(y)}{f_\theta(y)}\Big) - \Big(\frac{d(y)}{1-\tau} + \frac{f_\theta(y)}{\tau}\Big) \log\Big(\tau \frac{d(y)}{f_\theta(y)} + 1 - \tau\Big) \right]$$

$$G_{\mathrm{GKL}}(\delta_P) = \frac{\delta_P + 1}{1-\tau} \log(\delta_P + 1) - \frac{\tau\delta_P + 1}{\tau(1-\tau)} \log\Big(\tau\delta_P + 1\Big)$$

$$A_{\mathrm{GKL}}(\delta_P) = \frac{1}{\tau} \log\Big(\tau\delta_P + 1\Big)$$

Robust?	Inlier Robust?	Efficiency?
$\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	$\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	$A''(0)$
0	$-\frac{log(1-\tau)}{\tau}$	$-\tau$
Yes	No	Second Order when $\tau = 0$

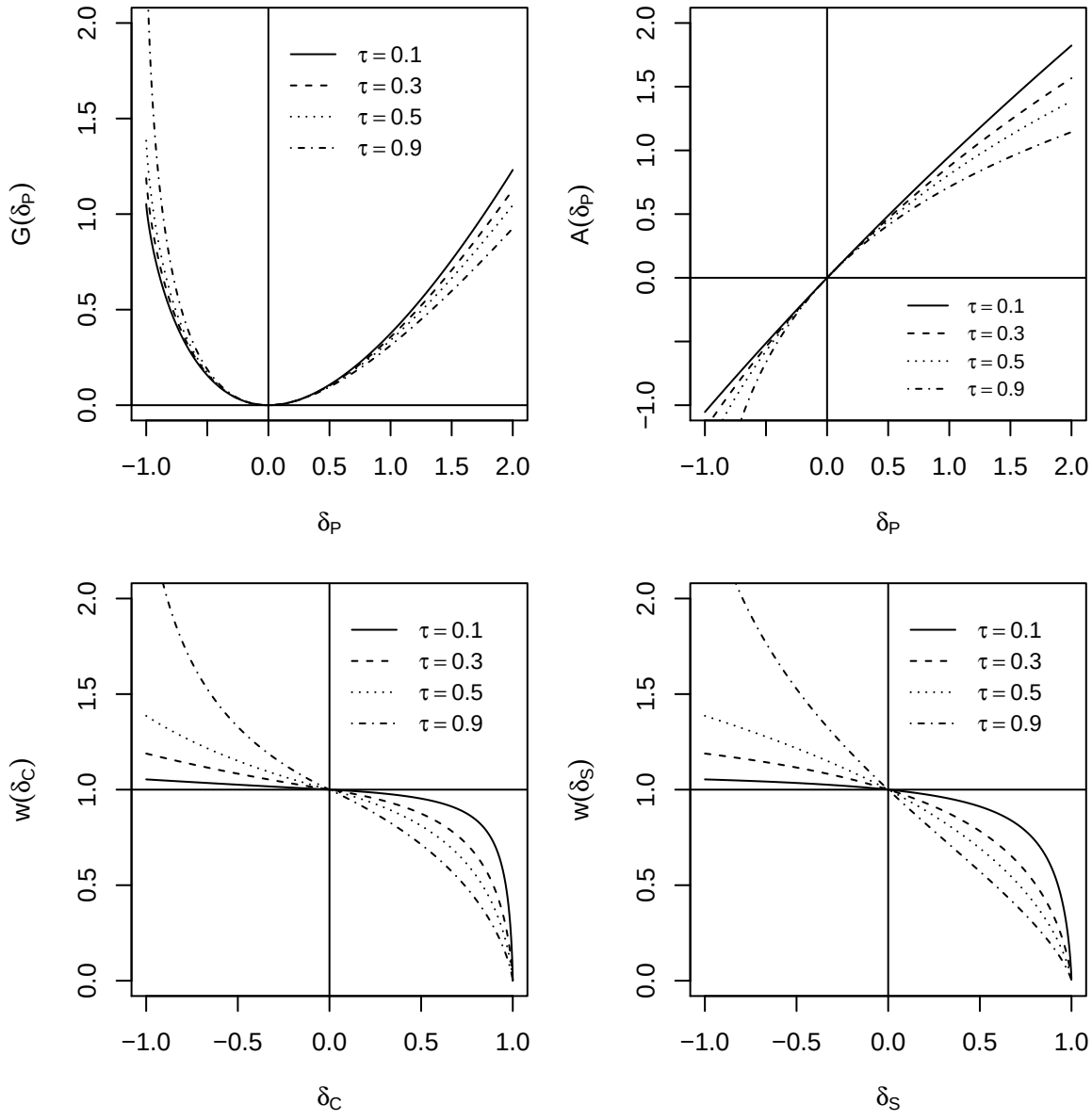

Figure A.7: Generalized Kullback-Leibler

8. Blended Weight Hellinger Distance Family (BWHD)

For $\tau \in [0,1]$ we have the following

$$\text{BWHD}_\tau(d(y), f_\theta(y)) = \frac{1}{2}\sum_{\forall y}\left[\frac{d(y) - f_\theta(y)}{\tau\sqrt{d(y)} + (1-\tau)\sqrt{f_\theta(y)}}\right]^2$$

$$G_{\text{BWHD}}(\delta_P) = \frac{\delta_P^2}{2\left[\tau\sqrt{\delta_P + 1} + 1 - \tau\right]^2}$$

$$A_{\text{BWHD}}(\delta_P) = \frac{\delta_P}{\left[\tau\sqrt{\delta_P + 1} + 1 - \tau\right]^2} + \frac{1-\tau}{2}\frac{\delta_P^2}{\left[\tau\sqrt{\delta_P + 1} + 1 - \tau\right]^3}$$

Robust?	Inlier Robust?	Efficiency?
$\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	$\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	$A''(0)$
0	$\frac{1}{2(\tau-1)^2}$	$1 - 3\tau$
Yes	No	First Order

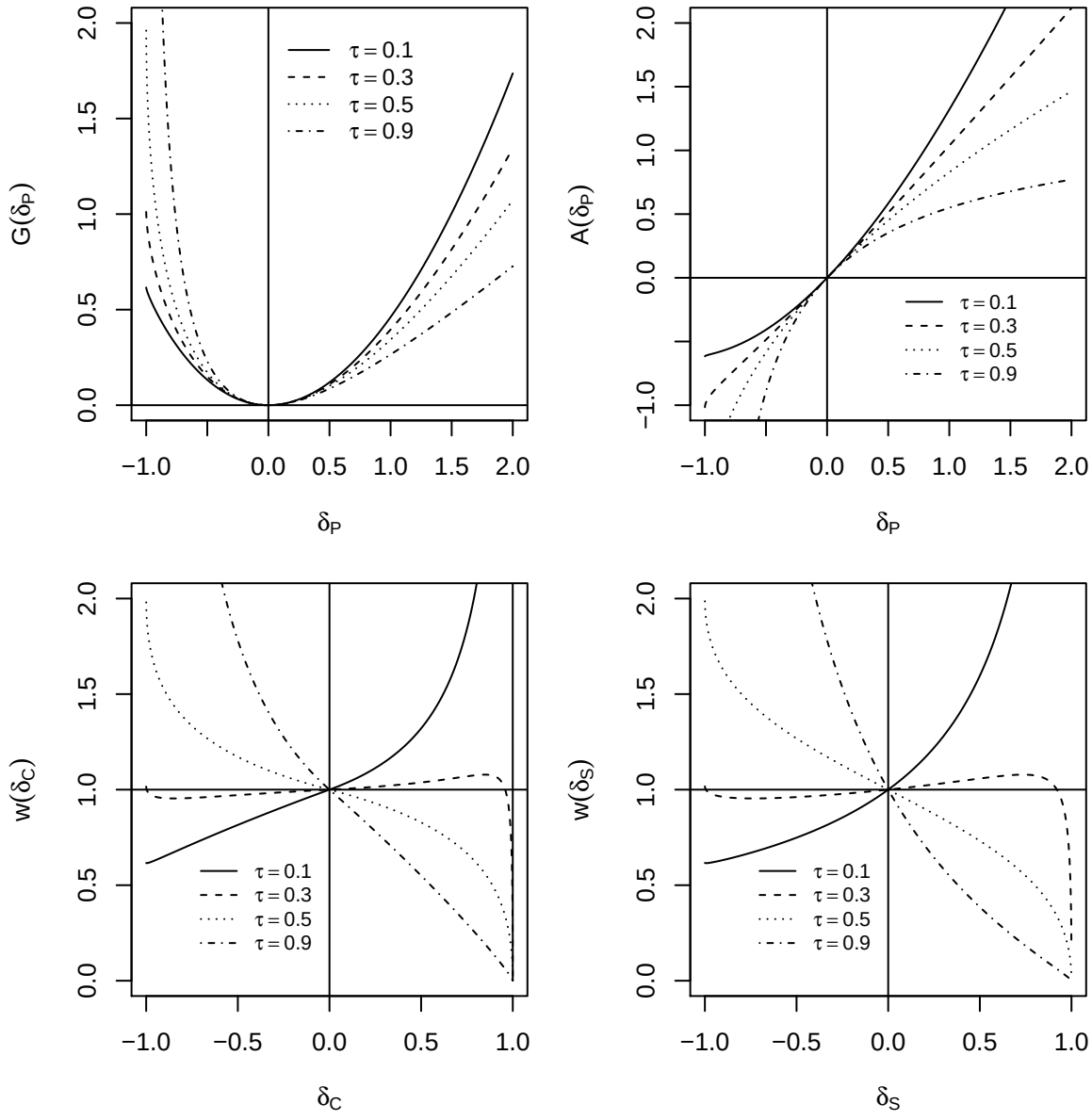

Figure A.8: Blended Weight Hellinger Distance

9. Blended Weight Chi-Square Distance Family (BWHD)

For $\tau \in [0,1]$ we have the following

$$\mathrm{BWCS}_\tau(d(y), f_\theta(y)) = \frac{1}{2}\sum_{\forall y} \frac{(d(y) - f_\theta(y))^2}{\tau d(y) + (1-\tau) f_\theta(y)}$$

$$G_{\mathrm{BWCS}}(\delta_P) = \frac{\delta_P^2}{2(\tau\delta_P + 1)}$$

$$A_{\mathrm{BWCS}}(\delta_P) = \frac{\delta_P}{\tau\delta_P + 1} + \frac{1-\tau}{2}\Big[\frac{\delta_P}{\tau\delta_P + 1}\Big]^2$$

Robust?	Inlier Robust?	Efficiency?
$\lim_{\delta_P \to \infty} \frac{A(\delta_P)}{\delta_P}$	$\lim_{\delta_P \to -1^+} \frac{A(\delta_P)}{\delta_P}$	$A''(0)$
0	$\frac{1}{2-2\tau}$	$1-3\tau$
Yes	No	First Order

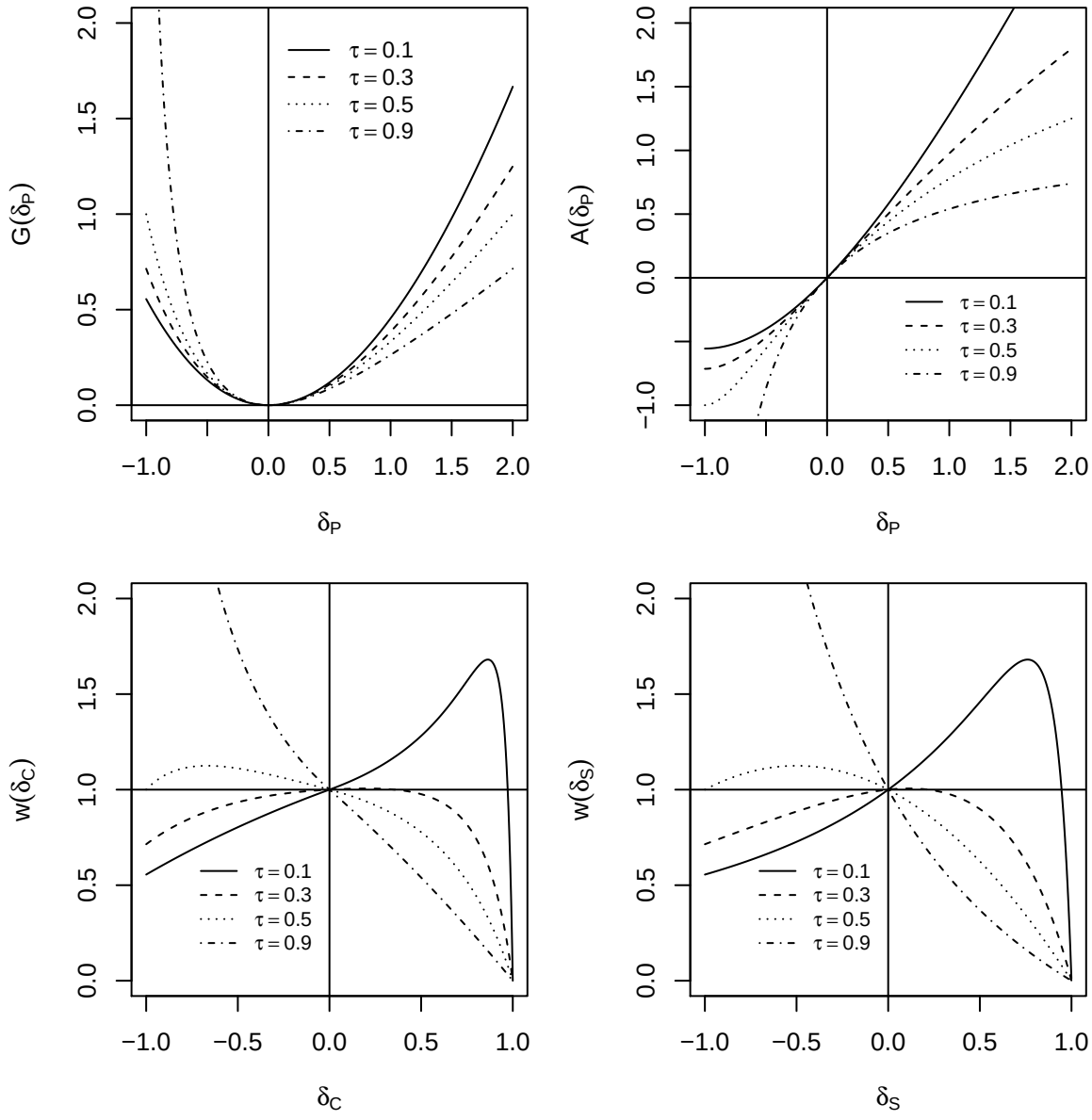

Figure A.9: Blended Weight Chi-Square

10. Symmetric Generalized Kullback-Leibler (SGKL)

For $\tau \in [0, 0.5]$ we have the following

$$\mathrm{SGKL}(d(y), f_\theta(y)) = \frac{\mathrm{GKL}_\tau(d(y), f_\theta(y)) + \mathrm{GKL}_\tau(f_\theta(y), d(y))}{2}$$

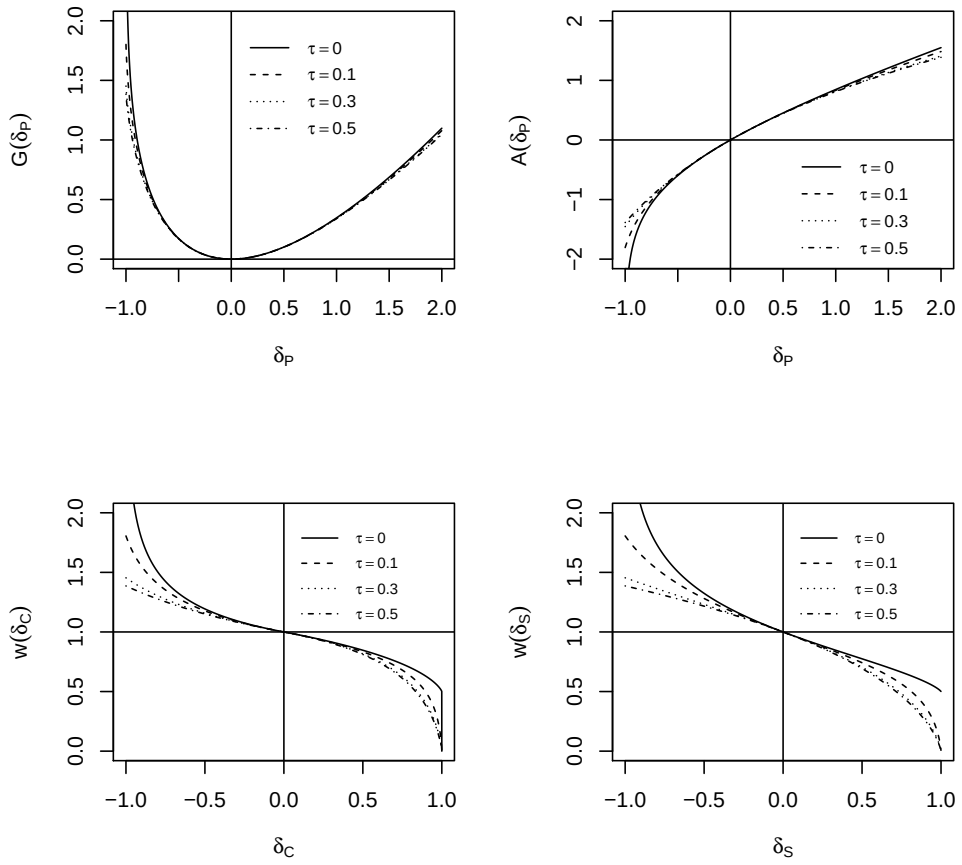

Figure A.10: Symmetric Generalized Kullback-Leibler

11. Symmetric Blended Weight Hellinger Distance Family (SBWHD)

For $\tau \in [0, 0.5]$ we have the following

$$\text{SBWHD}_\tau(d(y), f_\theta(y)) = \frac{\text{BWHD}_\tau(d(y), f_\theta(y)) + \text{BWHD}_\tau(f_\theta(y), d(y))}{2}$$

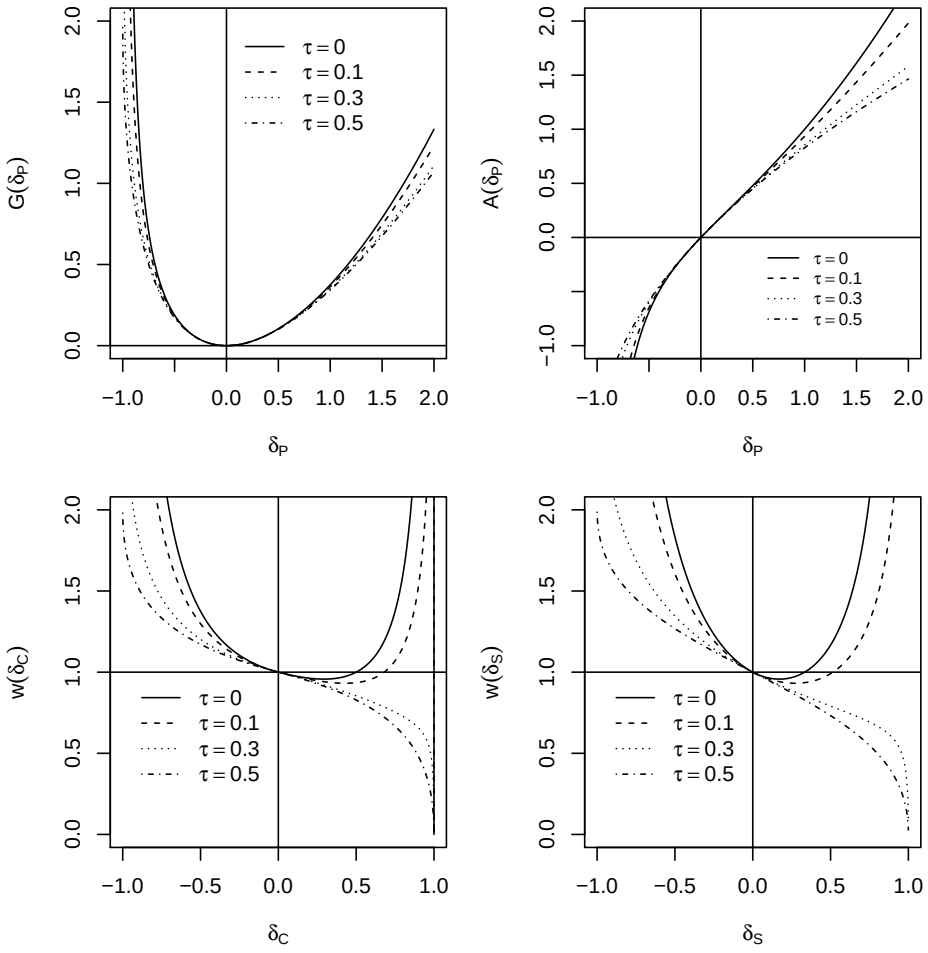

Figure A.11: Symmetric Blended Weight Hellinger Distance

12. Symmetric Blended Weight Chi-Square Distance Family (SBWHD)

For $\tau \in [0, 0.5]$ we have the following

$$\text{SBWCS}_\tau(d(y), f_\theta(y)) = \frac{\text{BWCS}_\tau(d(y), f_\theta(y)) + \text{BWCS}_\tau(f_\theta(y), d(y))}{2}$$

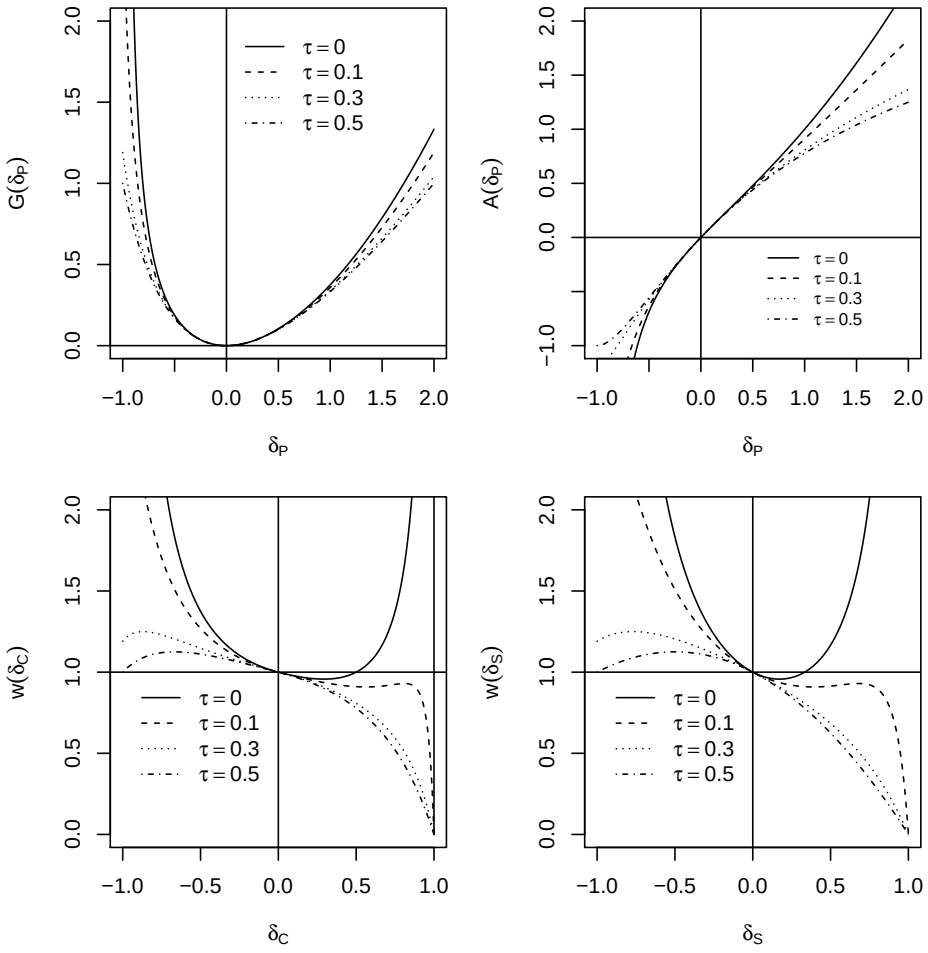

Figure A.12: Symmetric Blended Weight Chi-Square

Appendix B

Proving Upper Bounds

Proving Theorem 5.2.1

Proof. Let

$$\begin{aligned}\mathrm{BWHD}_\tau(d, f_\theta) &= \sum_{\forall y} \left[\frac{d(y) - f_\theta(y)}{\tau\sqrt{d(y)} + (1-\tau)\sqrt{f_\theta(y)}} \right]^2 \\ &= \sum_{\forall y} D(d(y), f_\theta(y))\end{aligned}$$

Differentiating the summand we get

$$\begin{aligned}\frac{\partial D}{\partial d} &= \frac{1}{2}\left[\frac{2(d(y) - f_\theta(y))}{(\tau\sqrt{d(y)} + (1-\tau)\sqrt{f_\theta(y)})^2} + \frac{-2(d(y) - f_\theta(y))^2}{(\tau\sqrt{d(y)} + (1-\tau)\sqrt{f_\theta(y)})^3} \cdot \left(\frac{\tau}{2\sqrt{d(y)}} \right) \right] \\ &= \frac{d(y) - f_\theta(y)}{(\tau\sqrt{d(y)} + (1-\tau)\sqrt{f_\theta(y)})^2} - \frac{\tau(d(y) - f_\theta(y))^2}{(\tau\sqrt{d(y)} + (1-\tau)\sqrt{f_\theta(y)})^3}\end{aligned}$$

Notice that $D(\cdot, f_\theta(y))$ is a strictly decreasing function on $0 < d(y) < f_\theta(y)$ and right continuous at $d(y) = 0$. Therefore, $D(d(y), f_\theta(y)) \leq D(0, f_\theta(y))$ for all $0 < d(y) < f_\theta(y)$ with equality only when $d(y) = 0$. Similarly, $D(d(y), f_\theta(y)) \leq D(d(y), 0)$ for all $0 < f_\theta(y) < d(y)$ with equality only when $f_\theta(y) = 0$. Therefore

$$
\begin{aligned}
\mathrm{BWHD}_\tau(d(y), f_\theta(y)) &= \sum_{\forall y} D(d(y), f_\theta(y)) \\
&= \sum_{d(y)\le f_\theta(y)} D(d(y), f_\theta(y)) + \sum_{d(y)> f_\theta(y)} D(d(y), f_\theta(y)) \\
&\le \sum_{\forall y} D(0, f_\theta(y)) + \sum_{\forall y} D(d(y), 0) \\
&= \frac{1}{2}\sum_{\forall y}\left(\frac{f_\theta(y)}{(1-\tau)\sqrt{f_\theta(y)}}\right)^2 + \frac{1}{2}\sum_{\forall y}\left(\frac{f_\theta(y)}{(1-\tau)\sqrt{f_\theta(y)}}\right)^2 \\
&= \frac{1}{2(1-\tau)^2}\sum_{\forall y} f_\theta(y) + \frac{1}{2\tau^2}\sum_{\forall y} d(y) \\
&= \frac{1}{2(1-\tau)^2} + \frac{1}{2\tau^2} \\
&= \frac{1}{2}\Big[\frac{1}{(1-\tau)^2} + \frac{1}{\tau^2}\Big]
\end{aligned}
$$

□

Proving Theorem 5.2.2

Proof. Let

$$
\begin{aligned}
\mathrm{BWCS}_\tau(d(y), f_\theta(y)) &= \frac{1}{2}\sum_{\forall y}\left[\frac{(d(y)-f_\theta(y))^2}{\tau d(y) + (1-\tau) f_\theta(y)}\right] \\
&= \sum_{\forall y} D(d(y), f_\theta(y))
\end{aligned}
$$

Differentiating the summand we get

$$
\begin{aligned}
\frac{\partial D}{\partial d} &= \frac{d(y)-f_\theta(y)}{\tau d(y) + (1-\tau) f_\theta(y)} + \frac{-(d(y)-f_\theta(y))^2}{\tau d(y) + (1-\tau) f_\theta(y)} \cdot \Big(\frac{\tau}{2}\Big) \\
&= \frac{d(y)-f_\theta(y)}{\tau d(y) + (1-\tau) f_\theta(y)} - \frac{\tau(d(y)-f_\theta(y))^2}{2\Big(\tau d(y) + (1-\tau) f_\theta(y)\Big)}
\end{aligned}
$$

Notice that $D(\cdot, f_\theta(y))$ is a strictly decreasing function on $0 < d(y) < f_\theta(y)$ and right continuous at $d(y) = 0$. Therefore, $D(d(y), f_\theta(y)) \leq D(0, f_\theta(y))$ for all $0 < d(y) < f_\theta(y)$ with equality only when $d(y) = 0$. Similarly, $D(d(y), f_\theta(y)) \leq D(d(y), 0)$ for all $0 < f_\theta(y) < d(y)$ with equality only when $f_\theta(y) = 0$. Therefore

$$
\begin{aligned}
\mathrm{BWCS}_\tau(d(y), f_\theta(y)) &= \sum_{\forall y} D(d(y), f_\theta(y)) \\
&= \sum_{d(y) \leq f_\theta(y)} D(d(y), f_\theta(y)) + \sum_{d(y) > f_\theta(y)} D(d(y), f_\theta(y)) \\
&\leq \sum_{\forall y} D(0, f_\theta(y)) + \sum_{\forall y} D(d(y), 0) \\
&= \frac{1}{2} \sum_{\forall y} \left(\frac{\Big(f_\theta(y)\Big)^2}{(1-\tau) f_\theta(y)} \right) + \frac{1}{2} \sum_{\forall y} \left(\frac{\Big(d(y)\Big)^2}{\tau d(y)} \right) \\
&= \frac{1}{2(1-\tau)} \sum_{\forall y} f_\theta(y) + \frac{1}{2\tau} \sum_{\forall y} d(y) \\
&= \frac{1}{2(1-\tau)} + \frac{1}{2\tau} \\
&= \frac{1}{2} \Big[\frac{1}{(1-\tau)} + \frac{1}{\tau} \Big] \\
&= \frac{1}{2\tau(1-\tau)}
\end{aligned}
$$

□

Bibliography

[Arifin and Asano, 2006] Arifin, A. and Asano, A. (2006). Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern Recognition Letters*, 27:1515–1521.

[Basseville, 2010] Basseville, M. (2010). Divergence measures for statistical data processing - an annotated bibliography. *Signal Processing*, 93:622–631.

[Basu et al., 2011] Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman and Hall CRC.

[Beran, 1977] Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, 5:445–463.

[Cressie and Read, 1984] Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. B*, 46:440–464.

[Fisher, 1922] Fisher, R. (1922). On the mathematical foundations of the theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368.

[Gonzalez and Woods, 2008] Gonzalez, R. and Woods, R. (2008). *Digital Image Processing*. Prentice Hall.

[He et al., 2003] He, Y., Hamza, A., and Krim, H. (2003). A generalized divergence measure for robust image registration. *IEEE*, 51:1211–1219.

[Jurekov and Sen, 1996] Jurekov, J. and Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations*. Wiley.

[Kindt and Coe, 2005] Kindt, R. and Coe, R. (2005). *Tree Diversity Analysis*. World Agroforestry Centre.

[Legendre and Gallagher, 2001] Legendre, P. and Gallagher, E. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129:271–280.

[Legendre and Legendre, 2012] Legendre, P. and Legendre, L. (2012). *Numerical Ecology*. Elsevier.

[Lindsay, 1994] Lindsay, B. (1994). Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The Annals of Statistics*, 22:1081–1114.

[Neyman, 1949] Neyman, J. (1949). Contribution to the theory of the chi-square test. *Berkeley: University of California Press*, pages 239–273.

[Park and Basu, 2003] Park, C. and Basu, A. (2003). The generalized kullback-leibler divergence and robust inference. *Journal of Statistical Computation and Simulation*, 73:311–332.

[Park et al., 2002] Park, C., Basu, A., and Lindsay, B. (2002). The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators. *Computational Statistics and Data Analysis*, 39:21–33.

[Pau et al.,] Pau, G., Oles, A., Smith, M., Skylar, O., and Huber, W. *EBImage: Image processing toolbox for R.* R package version 4.6.0.

[Pearson, 1900] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175.

[Rao, 1995] Rao, C. (1995). A review of canonical coordinates and an alternative to correspondence analysis using hellinger distance. *Questiio*, 19:23–63.

[Sandhu et al., 2008] Sandhu, R., Georgiou, T., and Tannenbaum, A. (2008). A new distribution metric for image segmentation. *Image Processing*, 6914.

[Schroff et al., 2006] Schroff, G. F., Criminisi, A., and Zisserman, A. (2006). Single-histogram class models for image segmentation. In *Proc. Indian Conference on Computer Vision*.

[Stigler, 2007] Stigler, S. M. (2007). The epic story of maximum likelihood. *Institute of Mathematical Statistics*, 22:598–620.

[V. Gonzalez-Castro and Alegre, 2013] V. Gonzalez-Castro, R. A.-R. and Alegre, E. (2013). Class distribution estimation based on the hellinger distance. *Information Sciences*, 218:146–164.

[Whittaker, 1972] Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, 21:213–251.